AD____________________

Award Number: DAMD17-03-1-0034

TITLE: Short- and Long-Term Effects in Prostate Cancer Survival: Analysis of Treatment Efficacy and Risk Prediction

PRINCIPAL INVESTIGATOR: Alexander Tsodikov, Ph.D.

CONTRACTING ORGANIZATION: Utah University
Salt Lake City, Utah 84102

REPORT DATE: March 2004

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 074-0188*

1. AGENCY USE ONLY ***(Leave blank)***	***2. REPORT DATE*** March 2004	***3. REPORT TYPE AND DATES COVERED*** Annual (1 Mar 2003 - 28 Feb 2004)

4. TITLE AND SUBTITLE Short- and Long-Term Effects in Prostate Cancer Survival: Analysis of Treatment Efficacy and Risk Prediction	***5. FUNDING NUMBERS*** DAMD17-03-1-0034
6. AUTHOR(S) Alexander Tsodikov, Ph.D.	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Utah University Salt Lake city, Utah 84102 *E-Mail:* atsodiko@hci.utah.edu	***8. PERFORMING ORGANIZATION REPORT NUMBER***
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012	***10. SPONSORING / MONITORING AGENCY REPORT NUMBER***

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited	***12b. DISTRIBUTION CODE***

13. ABSTRACT (Maximum 200 Words)

This report describes the work performed at the University of Utah during the first 8 months of the project. Algorithms have been developed for model for model building, estimation, hypothesis testing, and simulation for nonlinear transformation models of prostate cancer survival. These algorithms form the basis of a comprehensive toolbox for statistical inference with such models.

14. SUBJECT TERMS Survival effects in prostate cancer, statistical toolbox	***15. NUMBER OF PAGES*** 66
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	***18. SECURITY CLASSIFICATION OF THIS PAGE*** Unclassified	***19. SECURITY CLASSIFICATION OF ABSTRACT*** Unclassified	***20. LIMITATION OF ABSTRACT*** Unlimited

Contents

1 Statement of Work

Short- and Long-term Effects in Prostate Cancer Survival: Analysis of Treatment Efficacy and Risk Prediction
Alexander Tsodikov, Ph.D.

There has been no change in the statement of work. A breakdown below shows what has been accomplished in the first 8 months of the project at the University of Utah and the portion of the project to be completed at the University of California at Davis.

Tasks accomplished in the first 8 Months of the project at the University of Utah

Task 1. Develop model-building techniques (Months 1-4)

Task 2A. Develop estimation and hypothesis testing (Months 5-8)

(a) Develop point estimation
(b) Develop simulation algorithms
(c) Develop hypothesis testing

Tasks to be completed at the University of California at Davis

Task 2B. Develop estimation and hypothesis testing (Months 9-10)

(d) Develop software implementation
(e) Study models and methods by simulation

Task 3. Develop variable selection procedures (Months 11-13)

Task 4. Analyze data for significant effects (Months 14-18)

(a) Apply estimation, hypothesis testing and variable selection to MSKCC data and SEER data.
(b) Identify a model for prostate cancer biochemical recurrence, prostate cancer specific survival, and overall survival using methodology and software developed in Tasks 1-4.

Task 5. Computer-intensive approaches to prognosis and validation (Months 19-24)

Alex Tsodikov, Ph.D. Date: October 31, 2003.

2 Objectives

There has been no change in the project objectives. The specific aims of this project are

1. To provide a statistical model that reproduces the complex survival responses in prostate cancer.
2. To develop methodology for analysis of prognosis after treatment for prostate cancer taking into account the long- and short-term effects of prognostic factors and treatment.
3. To develop statistical software implementing model-building, estimation, construction of prognostic indices, conditional survival prognosis, and assessment of the quality of prognostic classifications based on the new models.
4. To apply the models and methodology to analyze post-treatment survival of patients with prostate cancer using data from the Memorial Sloan Kettering Cancer Center, the Utah Cancer Registry and the SEER database.

3 Introduction

The goal of this proposal is to investigate a novel approach to the analysis of post-treatment survival of prostate cancer patients: the decomposition of the diversity of survival patterns into short-term and long-term effects. We proposed to identify a model of prostate cancer survival incorporating long- and short-term effects of prognostic factors and treatment. Novel statistical tools are being developed to make such models work for better prognosis of prostate cancer patients. During the first 8 months of the project we focused on the development of algorithms for model building, estimation, hypothesis testing and simulation reviewed in the following sections of this progress report. Fragments of software implementation were also developed to illustrate the performance of the novel statistical tools using Surveillance Epidemiology and End Results (SEER) survival data by stage. With further developments in methodology and software planned in the project, data analysis will be refined to include a broad range of models, explanatory variables and variable selection techniques.

4 Estimation

4.1 Nonlinear Transformation Models (NTM)

4.1.1 PH mixture model

For a survival function $G(t \,|\, \boldsymbol{\beta}, \boldsymbol{z})$, where $\boldsymbol{\beta}$ are regression coefficients, and $\boldsymbol{z}$ are covariates, consider a PH mixture model

$$G(t \,|\, \boldsymbol{\beta}, \boldsymbol{z}) = \mathrm{E}\left\{ F(t)^{U(\boldsymbol{\beta}, \boldsymbol{z})} \,\middle|\, \boldsymbol{z} \right\}, \tag{1}$$

where F is the baseline survival function, and $U(\boldsymbol{\beta}, \boldsymbol{z})$ is a nonnegative random variable whose distribution depends on covariates and regression coefficients. This model can be considered a compact generalization of the so-called PH frailty model, or a PH model with a random effect

$$G(t \,|\, \boldsymbol{z}, \boldsymbol{z}) = \mathrm{E}\left\{ F(t)^{\theta(\boldsymbol{z})V} \right\}, \tag{2}$$

where θ is a predictor, and V is a random variable independent of the covariates, considered by Hougaard [1984], Klein [1992], Nielsen et al. [1992] and many other authors, for different distributions of V. Some authors considered specific frailty variables dependent on covariates, for example, Wassel and Moeschberger [1993], Clayton and Cuzick [1985]. Obviously, when V in (2) is allowed to depend on covariates, the model families (1) and (2) are equivalent.

We can make the following important observations about the class of PH mixture models (1):

- The survival function (1) is built by composition
$$G(t\,|\,\boldsymbol{\beta},\boldsymbol{z}) = (\gamma \circ F)(t\,|\,\boldsymbol{\beta},\boldsymbol{z}), \tag{3}$$
where $\gamma(x\,|\,\boldsymbol{\beta},\boldsymbol{z})$ is the moment generating function of U.

- The moment generating function $\gamma(x\,|\,\cdot)$ is a distribution function in x with the support on $[0,1]$. If the distribution of U is specified parametrically, γ is a parametric regression model on $[0,1]$.

- The fact that the range and the support of γ are the same allows one to build compositions of an arbitrary number of γs. One of the technical results we obtained is that the class of PH mixture models is closed with respect to such compositions.

These observations are used to generalize the PH mixture family into the NTM family (Section 4.1.2) and to develop the composition technique (Sections 5, 5.2) compatible with the QEM estimation procedure (Section 4.2).

We established the following key property of the PH mixture model [Tsodikov, 2003].

Proposition 4.1 *Suppose, we have an observation $(t,\boldsymbol{z},c)$ sampled from the PH mixture model under independent censoring, where t is an observed survival time and c is a censoring indicator ($c=0$ if t is a censored survival time, and $c=1$ if t is a failure). Then, under the PH mixture model (1),*

- *the conditional expectation of U, given the observed event $(t,\boldsymbol{z},c)$ is given by*
$$\mathrm{E}\left\{U(\cdot)\,|\,t,\cdot,c\right\} = (\Theta \circ F)(t\,|\,\cdot,c) = \Theta\left[F(t)\,|\,\cdot,c\right],$$
where the function Θ is given by
$$\Theta\left[x\,|\,\cdot,c\right] = c + x\frac{\gamma^{(c+1)}(x\,|\,\cdot)}{\gamma^{(c)}(x\,|\,\cdot)}, \tag{4}$$
where $\gamma^{(c)}(x\,|\,\cdot) = \partial^c\gamma(x\,|\,\cdot)/\partial x^c$, $c=0,1,\ldots$, $\gamma^{(0)}(x\,|\,\cdot)=\gamma(x\,|\,\cdot)$.

- *The function $\Theta\left[x\,|\,\cdot,c\right]$ is nondecreasing in x for any $c=0,1$.*

The nondecreasing character of the function Θ in the above statement is quite natural. The longer the subject stays event–free, the lower the subject's posterior risk, represented by Θ. So $\Theta\{F(t)\,|\,\cdot,c\}$ must be a nonincreasing function of t for both failure ($c=1$) and censoring ($c=0$) events. Since the survival function $F(t)$ is nonincreasing in t, $\Theta(x\,|\,\cdot,c)$ must be nondecreasing in x. It is interesting to note that the population hazard function for a heterogeneous population under the PH mixture model is expressed as $\lambda(t\,|\,\boldsymbol{z}) = \Theta\{F(t)\,|\,\cdot,0\}h(t)$, where h is the hazard function corresponding to F. Even if $h(t)$ is increasing, the observed population hazard function may be a decreasing one through the decreasing behavior of $\Theta\{F(t)|\cdot,0\}$ with time. This observation represents a selection effect of the risk set becoming "healthier" with time, as frail individuals leave the population. This effect was discovered and extensively studied in demography [Vaupel et al., 1979] in the context of misinterpretation of mortality trends.

4.1.2 Nonlinear Transformation Models

In Section 4.1.1 we considered semiparametric survival models of the form

$$G(t\,|\,\cdot) = \mathrm{E}\left\{F(t)^{U(\cdot)}\,\Big|\,\cdot\right\} = (\gamma \circ F)(t\,|\,\cdot),$$

where γ is a moment generating function of a nonnegative random variable U. We also noticed that $\gamma(x\,|\,\cdot)$ is a distribution function in $x \in [0,1]$ with the range contained in the same interval of $[0,1]$. This brings us to the following natural generalization of the PH mixture family of models.

Definition 4.1
Let $\gamma(x\,|\,\boldsymbol{\beta},\boldsymbol{z})$ be a parametrically specified distribution function with the x–domain of $[0,1]$. Let $F(t)$ be a nonparametrically specified baseline survival function. A semiparametric regression survival model is called a Nonlinear Transformation Model if its survival function can be represented in the form

$$G(t\,|\,\boldsymbol{\beta},\boldsymbol{z}) = \gamma\left\{F(t)\,|\,\boldsymbol{\beta},\boldsymbol{z}\right\} = (\gamma \circ F)(t\,|\,\boldsymbol{\beta},\boldsymbol{z}). \tag{5}$$

Functions γ will be called NTM–generating functions.

The class (5) was introduced in [Tsodikov, 2003], where universal estimation algorithms for the NTM class were developed (see Section 4.2). The key requirement that ensures monotonicity and convergence of the estimation algorithms of Section 4.2 is that of nondecreasing Θ, where Θ is defined in (4). Now that we no longer use the concept of frailty in the definition of NTM, $\Theta(F\,|\,\cdot,c)$ becomes a surrogate of the posterior risk such that its basic property of nondecreasing $\Theta(x\,|\,\cdot,c)$ is preserved. For these reasons, we will restrict the family of NTM–generating functions to those with nondecreasing Θ.

The class of NTM includes the class of linear transformation models [Cheng et al., 1995, 1997]

$$\log v(T|\boldsymbol{z}) = -\log\theta(\boldsymbol{z}) + \epsilon, \tag{6}$$

where T is the failure time, ϵ is the random error with the distribution μ, and v is some unspecified strictly increasing function. For the exponential predictor $\theta(\boldsymbol{\beta},\boldsymbol{z}) = \exp(\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z})$, the model assumes a linear form in covariates and transformed response. The connection between LTM, the PH model, the PO model, and binary regression models was discussed in [Doksum and Gasko, 1990]. After a little algebra, a linear transformation model can be represented as an NTM with the NTM–generating function

$$\gamma(x\,|\,\boldsymbol{\beta},\boldsymbol{z}) = p\left\{\log\theta(\boldsymbol{\beta},\boldsymbol{z}) + q(x)\right\}, \tag{7}$$

where p is a parametrically specified tail function (=1-distribution function), and q is an inverse tail function. It is convenient to specify q as the inverse of p, then $\theta = 1$ corresponds to the baseline $\gamma(x|\cdot) = x$.

Presentation of a semiparametric model in terms of an NTM–generating function γ is not unique, as there is a number of ways to represent an arbitrary monotonic function. Indeed, expression (7) suggests that a transformation $p\{q(F)\}$, where F is an arbitrary survival function, is again an arbitrary survival function. In other words, the family of functions

$$\tilde{\gamma}(x\,|\,\cdot) = (\gamma \circ p \circ q)(x\,|\,\cdot) \tag{8}$$

represents the same semiparametric model for any fixed p and q as defined above. The optimal choice of p and q to represent a semiparametric model in such a way as to achieve fastest convergence or to satisfy the assumption of nondecreasing Θ represents an interesting issue for future research.

It should be stressed that the issue of non-uniqueness of the representation of an NTM in terms of an NTM–generating function should not be confused with the issue of identiability of the model. For example, the Cox model can be represented as $\gamma(x|\theta(z)) = x^{\theta(z)}$ (proper form) or $\gamma(x|\theta(z)) = \exp\{-\theta(z)(1-x)\}$ (improper form, PH model with cure). In both models relative effects represented by regression coefficients entering the partial likelihood are identical and identifiable if predictor θ is coded correctly.

4.2 Quasi-EM Algorithm (QEM)

This section reviews a universal procedure (QEM algorithm) designed in [Tsodikov, 2003] to fit NTM models.

Let t_i, $i = 1, \dots, n$ be a set of times, arranged in increasing order, $t_{n+1} := \infty$. Associated with each t_i is a set of subjects $\mathcal{D}_i$ with covariates $\boldsymbol{z}_{ij}$, $j \in \mathcal{D}_i$ who fail at t_i, and a similar set of subjects $\mathcal{C}_i$ with covariates $\boldsymbol{z}_{ij}$, $j \in \mathcal{C}_i$ who are censored at t_i. The observed event $\mathcal{E}_{ij}$ for the subject ij is a triple $(t_i, \boldsymbol{z}_{ij}, c_{ij})$, where c is a censoring indicator, $c = 1$ if failure, $c = 0$ if right censored. For any function $A(t)$, let $A_i = A(t_i)$, $\Delta A_i = |A(t_i) - A(t_i - 0)|$. A step-wise function H can be characterized by two vectors $\Delta\boldsymbol{H} = (\Delta H_1, \dots, \Delta H_n)^{\mathrm{T}}$ and $\boldsymbol{t} = (t_1, \dots, t_n)^{\mathrm{T}}$. With this notation, under an NT model and noninformative censoring, the likelihood of survival data takes the form

$$\ell = \sum_{i=1}^{n} D_i \log[\Delta H_i] + \sum_{i=1}^{n} \sum_{j \in \mathcal{C}_i \cup \mathcal{D}_i} \log \vartheta(F_i \,|\, \boldsymbol{\beta}, \boldsymbol{z}_{ij}, c_{ij}), \tag{9}$$

where

$$\vartheta(x \,|\, \cdot, c) = x^c \gamma^{(c)}\{x \,|\, \cdot\},$$

and D_i is the number of failures associated with t_i. We use the profile likelihood approach to maximize ℓ. The profile likelihood is defined as a supremum of the full likelihood taken over the nonparametric part of the model

$$\ell_{pr}(\boldsymbol{\beta}) = \max_{H} \ell(\boldsymbol{\beta}, H). \tag{10}$$

The algorithm follows the straightforward nested procedure:

- Maximize $\ell_{pr}(\boldsymbol{\beta})$ by a conventional nonlinear programming method, for example, the Powell method [Press et al., 1994].

- For any $\boldsymbol{\beta}$ as demanded in the above maximization procedure, solve the problem (10).

Inference based on the profile likelihood is not straightforward, as the usual theory of MLE does not apply to unlimited dimension. Important results have been obtained regarding theoretical justification for the nonparametric maximum likelihood estimation (NPMLE) method and the profile likelihood for semiparametric models [Murphy, 2000, van der Vaart, 1998, Murphy and van der Vaart, 1997]. It was shown that profile likelihoods with nuisance estimated out behave like ordinary likelihoods under some conditions. In particular, these results apply to the PH model, the proportional odds (PO) model [Murphy, 2000, Murphy et al., 1997] and the PH frailty model [Murphy, 1994, 1995], and presumably to most other models.

The following method (QEM) is used to obtain the profile likelihood and solve (10):

$$\Delta H_m^{(k+1)} = \frac{D_m}{\sum_{ij \in \mathcal{R}_m} \Theta(F_i^{(k)} \,|\, \boldsymbol{\beta}_{ij}, \boldsymbol{z}_{ij}, c_{ij})}, \tag{11}$$

where k counts iterations.

It can be shown that if Θ is nondecreasing, each update of H using (11) strictly improves the likelihood, given $\boldsymbol{\beta}$. This guarantees convergence of the sequence of likelihood values $\ell\left\{\boldsymbol{\beta}, H^{(k)}\right\}$ to the profile likelihood under fairly general conditions.

Under a PH mixture model, the procedure (11) is an EM algorithm based on imputation of the missing predictor U in the Nelson–Aalen–Estimator by its conditional expectation, given observed data, represented by $\Theta(F \,|\, \boldsymbol{\beta}, \boldsymbol{z}, c)$. Under an NT model, the procedure works as a Quasi-EM algorithm without the missing-data interpretation. It can be shown [Tsodikov, 2003] that imputation in the QEM procedure is accomplished using the so-called quasi–expectation operator, QE, which generalizes mathematical expectation operator on a restricted class of basis functions in a way that its linearity and second-order differentiation properties, as well as the Jensen inequality are preserved.

5 Model Building

5.1 Composition for PH mixture models

The idea to use compounding to build particular extended families of frailty models is not new. For example, Aalen [1992] used a compound Poisson distribution to extend a class of frailty models by Hougaard [1984].

Consider the following general compounding techniques for the PH mixture model. If ν is a nonnegative discrete random variable with the moment generating function $\gamma_\theta(x) = \mathrm{E}\left\{x^\nu\right\}$, and ξ_k are i.i.d. copies of another nonnegative random variable (independent of ν) with the the moment generating function $\gamma_\eta(x) = \mathrm{E}\left\{x^\xi\right\}$, and U is a compound random variable given by

$$U = \sum_{k=1}^{\nu} \xi_k, \tag{12}$$

then by the composition property of Laplace transform,

$$\gamma(x) = \mathrm{E}\left\{x^U\right\} = (\gamma_\theta \circ \gamma_\eta)(x). \tag{13}$$

A large variety of semiparametric mixture models can be derived from (13). When $\gamma_\theta(x)$ corresponds to a continuous random variable, the compound variable U is no longer of the simple form (12). However, the composition $\gamma_\theta \circ \gamma_\eta$ still corresponds to a PH mixture model with some frailty random variable U, as given by the following proposition.

Proposition 5.1 *Composition for mixture models.*
Let γ_θ and γ_η be some two mixture models $\gamma_\theta(x|\cdot) = E(x^\nu \,|\, \cdot)$, $\gamma_\eta(x|\cdot) = E(x^\xi \,|\, \cdot)$, where ν and ξ are some independent nonnegative random variables. Let $\gamma = \gamma_\theta \circ \gamma_\eta$ be the compound model. Then γ is also a mixture model, meaning that there exists a nonnegative random variable U such that $\gamma(x|\cdot) = E(x^U \,|\, \cdot)$.
Proof. By the Bernstein theorem (Feller [1971]), we need to prove that $\gamma(e^{-s}|\cdot) = (\gamma_\theta \circ \gamma_\eta)(e^{-s}|\cdot)$ is a completely monotonic function. Let $\psi_\cdot(s) = \gamma_\cdot(e^{-s}|\cdot)$. We have $\psi(s) = \psi_\theta\left\{-\log \psi_\eta(s)\right\}$. For any functions ξ and ζ, the composition $\xi \circ \zeta$ is completely monotonic if ξ is completely monotonic, $\zeta > 0$, and ζ' is completely monotonic. Applied to the functions ψ, this means that we have to prove that for any completely monotonic function $\psi(s) > 0$, the function $f(s) = \{-\log \psi(s)\}'$ is

completely monotonic. It can be proved by induction that

$$(-1)^n f^{(n)}(s) = \sum_{k=1}^{n+1} a_{nk}(-1)^k \frac{\psi^{(k)}(s)}{\psi^{n-k+2}(s)},$$

where $a_{01} = 1$, $a_{n+1,1} = a_{n1}$, $a_{n+1,k} = a_{nk}(n-k+2) + a_{n,k-1}$, $k = 2, \ldots, n+1$, $a_{n+1,n+2} = a_{n,n+1}$, $n = 0, 1, \cdots$. From the above equations it follows that $a_{nk} > 0$ for any n, k. Also, $\psi(s) > 0$, $s > 0$, and since ψ is completely monotonic, $(-1)^k \psi^{(k)}(s) \geq 0$. Therefore, $(-1)^n f^{(n)}(s) \geq 0$, $s > 0$. End of proof.

As a result of the above observations, we have a tool to build hierarchical regression models using composition of moment generating functions $\gamma = \gamma_\theta \circ \gamma_\eta$. The fact that the class of PH mixture models is closed with respect to such composition allows us to use an EM-approach to fit compound mixture models.

Consider a model composed of the PH and the PO models. Take the moment generating function

$$\gamma_\eta(x| \cdot) = \frac{\eta(\cdot)}{\eta(\cdot) - \log x}$$

of the exponential distribution with parameter $\eta(\boldsymbol{\beta}, \boldsymbol{z})$, corresponding to the PO model. Take another moment generating function

$$\gamma_\theta(x| \cdot) = x^{\theta(\cdot)},$$

corresponding to the PH model with predictor $\theta(\boldsymbol{\beta}, \boldsymbol{z})$. As a result of the composition $\gamma = \gamma_\theta \circ \gamma_\eta$, we have the so-called Γ–frailty model

$$G\{t| \theta(\cdot), \eta(\cdot)\} = \left\{\frac{\eta(\cdot)}{\eta(\cdot) + H(t)}\right\}^{\theta(\cdot)}. \tag{14}$$

Indeed, $\psi(s) = \gamma(e^{-s}|\cdot) = [\eta(\cdot)/\{\eta(\cdot) + s\}]^{\theta(\cdot)}$ is the Laplace transform of a Γ-distribution with scale parameter η and shape parameter θ, and we have the interpretation of the compound model (14) as a Γ–frailty model.

It is assumed that predictors depend on $\boldsymbol{\beta}, \boldsymbol{z}$ via the form $\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}$, where β_0 stands for the intercept term of the predictor. Also, different predictors have independent sets of regression coefficients $\theta = \theta(\beta_{10} + \boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{z})$, $\eta = \eta(\beta_{20} + \boldsymbol{\beta}_2^{\mathrm{T}}\boldsymbol{z})$. To avoid overparameterization of the Γ-frailty model, the intercept in θ is fixed at zero $\beta_{10} = 0$. With the above conventions, a test for $\boldsymbol{\beta}_1 = \boldsymbol{0}$ is a test for the PO assumption, while a test for $\boldsymbol{\beta}_2 = \boldsymbol{0}$ is a test for the PH assumption. Setting $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{0}$ corresponds to a test of homogeneity.

5.2 Composition for NTMs

We extended the composition techniques for the PH mixture model (Section 5.1) to the NTM class. The composition technique offers a simple way to build hierarchical families of models that combine the features of simpler models. Specifically, if γ_θ and γ_η are two different NT models with predictors θ, and η, respectively, then

$$\gamma(x|\cdot) = (\gamma_\theta \circ \gamma_\eta)(x|\cdot) \tag{15}$$

is a new semiparametric model with two predictors θ and η. If $\gamma_\theta(x|\cdot) \equiv x$ for some value of θ (usually for $\theta = 1$), then the model (15) includes models γ_θ and γ_η as nested special cases. The fact that NTM–generating functions $\gamma(x \mid \cdot)$ are all defined on $x \in [0, 1]$ and have the range in the same interval allows us to compose as complex a hierarchical model as needed. In particular, a composition of the PH model in the improper form

$$\gamma(x|\theta(z)) = \exp\{-\theta(z)(1-x)\} \tag{16}$$

with the one in the proper form

$$\gamma_\eta(x|\eta(z)) = x^{\eta(z)} \tag{17}$$

results in the so-called PHPH model incorporating long- and short-term survival effects

$$G(t|\boldsymbol{z}) = \exp\left[-\theta(\boldsymbol{z})\left\{1 - [F(t)]^{\eta}(\boldsymbol{z})\right\}\right]. \tag{18}$$

We have shown that operation of composition preserves the key property of nondecreasing Θ, and the EM–like estimation algorithms of Section 4.2 remain applicable within the hierarchical family.

Proposition 5.2 *Composition.*
Let γ_θ and γ_η be some two NTM–generating functions, each satisfying the assumption of nondecreasing Θ, where Θ is given by (4), and let $\gamma = \gamma_\theta \circ \gamma_\eta$ be the compound function (compositions are taken with respect to x). Let Θ_a be the Θ–function (4) corresponding to γ_a, $a = \theta, \eta$, and to the compound function γ, if a is blank. Then (A)

$$\Theta(x \mid \cdot, c) = \Theta_\eta(x \mid \cdot, 0)\left\{(\Theta_\theta \circ \gamma_\eta)(x \mid \cdot, c) - c\right\} + c\Theta_\eta(x \mid \cdot, c), \tag{19}$$

where $c = 0, 1$ and $(\Theta \circ \gamma)(x \mid \cdot, c)$ is understood as $\Theta\{\gamma(x \mid \cdot)|\cdot, c\}$; and
(B) The function Θ (4) derived from the compound NTM–generating function γ is nondecreasing in x as required for monotonicity and convergence of the estimation algorithms (See Section 4.2).
Proof. Proof of first statement is a straightforward exercise in differentiation of compound functions entering (4). Validity of second statement follows from (19) upon observation that all components of (19) are nondecreasing functions in x. End of proof.

Equation (19) simplifies derivation of Θ for the compound models through direct use of Θs corresponding to submodels participating in the composition.

Consider another example of building hierarchical models using composition. Using the composition framework, the Dabrowska and Doksum model [Dabrowska and Doksum, 1988] can be represented through a composition $\gamma = \gamma_{1/a} \circ \gamma_\theta \circ \gamma_a$, where γ_θ is an NTM–generating function for the PO model, γ_a and $\gamma_{1/a}$ correspond to the PH model, and a is a scalar, independent of covariates

$$\gamma(x \mid \cdot) = \left\{\frac{\theta(\cdot)x^a}{1 - \bar{\theta}(\cdot)x^a}\right\}^{\frac{1}{a}}, \quad a \geq 0. \tag{20}$$

The above model becomes the PO model if $a = 1$, and it becomes the PH model in the limit as $a \to 0$. With the above model, the PH assumption corresponds to the border of the parametric space ($a = 0$) and, for this reason, we prefer to use the Γ–frailty model (14) in this paper to illustrate the methodology.

The Γ–frailty model (14) can be built as a composition of the NTM–generating functions

$$\gamma_\theta(x| \cdot) = x^{\theta(\cdot)}$$

and

$$\gamma_\eta(x| \cdot) = \frac{\eta(\cdot)}{\eta(\cdot) - \log x}$$

corresponding to the PH and the PO models, respectively, without having to work with their frailty underpinnings. Using (4), we derive the Θ–functions corresponding to the two submodels:

- PH model: $\Theta_\theta(x \mid \cdot, c) \equiv \theta(\cdot)$,

- PO model: $\Theta_\eta(x \,|\, \cdot, c) = (c+1)\{\eta(\cdot) - \log x\}^{-1}$.

Next, the compound Θ is derived from (19) using the above expressions for the submodels:

$$\Theta(x \,|\, \cdot, c) = \frac{\theta(\cdot) + c}{\eta(\cdot) - \log x}. \tag{21}$$

6 Hypotheses Testing

It should be noted that the problem of deriving a variance estimator with the QEM procedure is quite different from the one with the EM algorithm. The problem with the EM-based information matrix is that the likelihood is unavailable in a closed form, or it is difficult to differentiate. In our situation, however, the problem is formulated without use of missing data, and the likelihood is a quite simple, easily differentiable function. So, it is not a problem to write down the full model information matrix. The problem is that the number of parameters of a semiparametric model is potentially unlimited. For this reason, obtaining the inverse of the full information matrix can be computationally prohibitive.

We have developed a numerically efficient procedure to overcome the problem using the profile information matrix

$$\boldsymbol{I}^{\mathrm{P}}_{\boldsymbol{\beta},\boldsymbol{\beta}} = -\frac{\partial^2 \ell_{pr}\{\boldsymbol{\beta}\}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}, \tag{22}$$

where ℓ_{pr} is the profile likelihood (10) obtained by searching for the fixed point $\Delta\boldsymbol{H}^*(\boldsymbol{\beta})$ of (11). Implicit differentiation of the profile likelihood yields

$$\boldsymbol{I}^{\mathrm{P}}_{\boldsymbol{\beta},\boldsymbol{\beta}} = \boldsymbol{I}_{\boldsymbol{\beta},\boldsymbol{\beta}} + \left(\frac{\partial\Delta\boldsymbol{H}^*}{\partial\boldsymbol{\beta}}\right)^{\mathrm{T}} \boldsymbol{I}_{\Delta\boldsymbol{H},\Delta\boldsymbol{H}} \frac{\partial\Delta\boldsymbol{H}^*}{\partial\boldsymbol{\beta}} + \left(\frac{\partial\Delta\boldsymbol{H}^*}{\partial\boldsymbol{\beta}}\right)^{\mathrm{T}} \boldsymbol{I}_{\Delta\boldsymbol{H},\boldsymbol{\beta}} + \boldsymbol{I}^{\mathrm{T}}_{\Delta\boldsymbol{H},\boldsymbol{\beta}} \frac{\partial\Delta\boldsymbol{H}^*}{\partial\boldsymbol{\beta}}, \tag{23}$$

where

$$\boldsymbol{I}_{\boldsymbol{a},\boldsymbol{b}} = -\frac{\partial^2\ell}{\partial\boldsymbol{a}\partial\boldsymbol{b}^{\mathrm{T}}}$$

for any two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. To get a variance estimator for regression coefficients $\boldsymbol{\beta}$, only a small ($\dim\boldsymbol{\beta} \times \dim\boldsymbol{\beta}$) profile matrix needs to be inverted. The downside of (23) is that the Jacobian matrix $\partial\Delta\boldsymbol{H}^*/\partial\boldsymbol{\beta}$ is generally unavailable in a closed form. The success of the above approach is determined by the existence of an efficient numerical method to compute $\partial\Delta\boldsymbol{H}^*/\partial\boldsymbol{\beta}$. Generally, computation of $\partial\Delta\boldsymbol{H}^*/\partial\boldsymbol{\beta}$ is as difficult as taking the inverse of the original full model information matrix ($O(n^3)$ operations required), and this derivation defeats the purpose.

However, if the functional $\vartheta(H, t \,|\, \cdot)$ depends on (H, t) only through $H(t)$, which corresponds to the NT models (Section 4.1.2), the system of linear equations for $\partial\Delta\boldsymbol{H}^*/\partial\boldsymbol{\beta}$ acquires a triangular form, and it can be solved recurrently. Indeed, with an NT model we have

$$\frac{\partial\ell}{\partial\Delta H_k} = \frac{D_k}{\Delta H_k} - \sum_{ij\in\mathcal{R}_k} \Theta(F_i|\boldsymbol{\beta}, \boldsymbol{z}_{ij}, c_{ij}), \tag{24}$$

where Θ is given by (4). Differentiating the above score equation with respect to ΔH_m, we get the elements of the $\boldsymbol{I}_{\Delta\boldsymbol{H},\Delta\boldsymbol{H}}$ information matrix

$$-\frac{\partial^2\ell}{\partial\Delta H_k\partial\Delta H_m} = \frac{D_k}{\Delta H_k^2}\delta_{km} + \sum_{ij\in\mathcal{R}_{\max(k,m)}} Q(F_i|\boldsymbol{\beta}, \boldsymbol{z}_{ij}, c_{ij}),$$

where $\delta_{km} = 1$, if $k = m$ and $\delta_{km} = 0$ otherwise, and

$$Q(x|\cdot, c) = -\frac{\partial\Theta(x|\cdot)}{\partial\log(x)} = \{\Theta(x|\cdot, c) - c\} \times \{\Theta(x|\cdot, c) - \Theta(x|\cdot, c+1)\}. \tag{25}$$

Consider the self consistency equation (11) at the fixed point $\Delta\boldsymbol{H}^*(\boldsymbol{\beta})$. Implicit differentiation of the likelihood after a little algebra gives the following system of equations

$$\frac{\partial H^*_{k-1}}{\partial\boldsymbol{\beta}} = C_k\left(\frac{\partial H^*_k}{\partial\boldsymbol{\beta}} + \boldsymbol{A}_k + \sum_{i=k}^{n} B_i\frac{\partial H^*_i}{\partial\boldsymbol{\beta}}\right), \tag{26}$$

where

$$\boldsymbol{A}_k = \sum_{ij\in\mathcal{R}_k}\frac{\partial\Theta(F^*_i|\boldsymbol{\beta}, \boldsymbol{z}_{ij}, c_{ij})}{\partial\boldsymbol{\beta}},$$
$$B_i = \sum_{j\in\mathcal{C}_i\cup\mathcal{D}_i} Q(F^*_i|\boldsymbol{\beta}, \boldsymbol{z}_{ij}, c_{ij}),$$

and

$$C_k = \frac{(\Delta H^*_k)^2}{D_k}.$$

The system of equations (26) is linear in $\partial H^*_k/\partial\boldsymbol{\beta}$ with an upper triangular matrix. Such systems can be solved recurrently. On substitution of the above solutions into (23), the profile information matrix for an NT model is obtained.

We are working on software implementation and numerical experiments with the above method.

Also, we have implemented a pragmatic numerical approach similar to the one proposed in reference [Nielsen et al., 1992]. In the course of maximization of the profile likelihood with respect to regression coefficients $\boldsymbol{\beta}$, a dense sample of the profile likelihood surface is generated in the vicinity of a stationary point. The curvature of the profile likelihood surface at the stationary point can be estimated by fitting a quadratic function to some domain around the point, using least squares. For example, the domain can be limited to points that cannot be rejected using the likelihood ratio test (applied informally). Our numerical experiments have shown that this method is unstable for models with more than one or two covariates. Also, it is much less efficient computationally compared to the one based on implicit differentiation of the profile likelihood.

7 Preliminary Data Analysis

As an example, we used data from the National Cancer Institutes Surveillance Epidemiology and End Results (SEER) program. Using the publicly available SEER database, 11621 cases of primary prostate cancer diagnosed in the state of Utah between 1988 and 1999 were identified. The following selection criteria were applied to a total of 19819 Utah–cases registered in the database: valid positive survival time, valid stage of the disease, age$\geq$ 18 years. Prostate cancer specific survival was analyzed by stage of the disease (localized/regional vs. distant). For the definition of stages as well as for other details of the data we refer the reader to SEER documentation http://seer.cancer.gov/.

Three models PH, PO, and Γ–frailty model with $\boldsymbol{z}$ representing two groups corresponding to localized/regional stage (10765 cases) and distant stage (856 cases), respectively, were fitted using the profile QEM algorithm. Estimates of model parameters are given in Table 1.

Observed (Kaplan–Meier) and expected model–based estimates of the survival functions by group as well as diagnostic plots are shown in Figure 1.

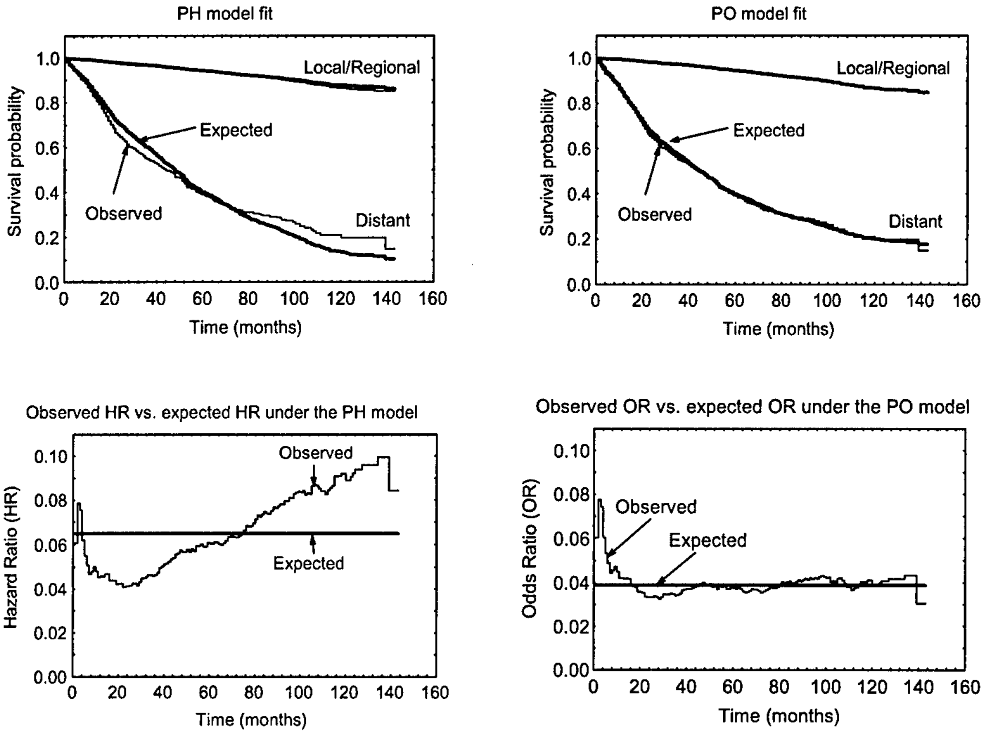

Figure 1: Observed vs. expected plots corresponding to the PH and the PO model fitted to the prostate cancer data. Thin and thick lines correspond to observed and expected plots, respectively.

It is evident from the comparison of observed and expected survival functions that the PO model provides better fit to the data than the PH model. The Γ–frailty model provides the fit (not shown) which is very similar to the PO model. We test the model assumptions using the hierarchical structure of the Γ–frailty model. Using the likelihood ratio test with the profile likelihood, the PO assumption could not be rejected (p=0.712).

8 Key Research Accomplishments

Simmarizing, the key research accomplishments of the first 8 months period of the project are:

1. Development of the class of Nonlinear Transformation Models (NTM) and associated QEM estimation procedures and their computer implementation;

2. Development of composition technique as a tool for model building. We established that the class of PH mixture models and NTM is closed with respect to compositions of so-called model generating functions. Moreover, we established that QEM estimation procedures are applicable to any model built using this technique;

3. A numerically efficient algorithm has been developed for estimation of the inverse of profile information matrix for NTM that avoids taking inverse of the full information matrix of the semiparametric model. This algorithm will be used for testing hypotheses, variable selection and estimation of confidence intervals.

Model	Parameter	Point– estimate	Confidence interval	p-Value	Test
PH	β_{PH}	2.734	(2.611,2.858)	<0.0001	Homogeneity
PO	β_{PO}	-3.251	(-3.416,-3.086)	<0.0001	Homogeneity
Γ–frailty	β_{PH}	0.071	(-0.310,0.452)	0.712	$\beta_{\mathrm{PH}} = 0$, PO
	β_{PO}	-3.149	(-3.714,-2.583)	<0.0001	$\beta_{\mathrm{PO}} = 0$, PH
	–	–	–	<0.0001	Homogeneity

Table 1: Parameter estimation and hypothesis testing for prostate cancer data based on PH, PO, and Γ–frailty models. Regression coefficients β_{PH} and β_{PO} measure the disadvantage of being in the distant stage relative to local/regional stage as represented by log odds ratio and log hazards ratio, respectively. Negative β_{PO} and positive β_{PH} means worse survival.

9 Reportable Outcomes

9.1 Manuscripts

1. Tsodikov, A. (2003) Semiparametric models: A generalized self-consistency approach, *Journal of the Royal Statistical Society*, Series B, Vol. 65, 759-774.

2. Tsodikov, A., Ibrahim, J.G., and Yakovlev, A.Y. (2003) Estimating Cure Rates from Survival Data: An Alternative to Two-Component Mixture Models, *Journal of the American Statistical* Association, (review paper) to appear in December 2003.

3. Boucher, K., Asselain, B., Tsodikov, A., Yakovlev, A. Semiparametric versus parametric regression analysis based on the Bounded Cumulative Hazard Model: An application to breast cancer recurrence, In Semiparametric Models in Survival Analysis, Quality of Life and Reliability, Birkhauser (invited paper), to appear.

9.2 Presentations

Tsodikov, A. (2003) Generalized Self-Consistency Methods for Cure Models, Joint Statistical Meetings, Invited session on Cure Models. (invited), San Francisco, August 2003.

10 Conclusions

Most semiparametric survival models can be induced by frailties. Compounding the distribution of frailty offers a way to build hierarchical families of semiparametric models that can be used to test model assumptions and to reproduce complex patterns of covariate effects using more than one predictor. A PH mixture model represents survival function G as a composition $G = \gamma \circ F$, where γ is a moment generating function, and F is a nonparametrically specified baseline survival function. We note that hierarchical PH mixture models can be built using compositions of the form $G = \gamma_1 \circ \ldots \circ \gamma_m \circ F$, where γ_i are moment generating functions for submodels. Mixture models can be fit by an EM algorithm which is specified as repeated imputation of the missing frailty variable using its conditional expectation, given observed event. We find that this conditional expectation is represented as a composition $\Theta \circ F$, where Θ is defined through first two derivatives of γ. The above framework is naturally generalized as the existence of frailty interpretation is required neither for model building nor for model fitting. For the class of Nonlinear Transformation Models, we formulate composition rules for γ and Θ and develop the QEM generalization of the EM algorithm.

During the first phase of the project we developed a basic toolbox of analytic and algorithmic tools for model building, model fitting and hypothesis testing. Fragments of software implementation were developed and used to demonstrate the utility and superior numerical performance of these methods in real and simulated data.

During the second phase of the project we are planning to continue to build on the results obtained so far extending the arsenal of methods to include variable selection and computer-intensive methods. We will continue software implementation of these procedures under a common shell, verification of the methods by simulations and their application to real prostate cancer data.

11 References

References

O.O. Aalen. Modelling heterogeneity in survival analysis by the compound Poisson distribution. *Annals of Applied Probability*, 2:951–972, 1992.

S.R. Cheng, L.J. Wei, and Z Ying. Analysis of transformation models with censored data. *Biometrika*, 82:835–845, 1995.

S.R. Cheng, L.J. Wei, and Z Ying. Predicting survival probabilities with semiparametric transformation models. *Journal of the American Statistical Association*, 92(437):227–235, 1997.

D. Clayton and J. Cuzick. Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society, Series A*, 148:82–117, 1985.

D.M. Dabrowska and K.A. Doksum. Estimation and testing in a two-sample generalized odds-rate model. *Journal of the Americal Statistical Association*, 83:744–749, 1988.

K.H. Doksum and M. Gasko. On a correspondence between models in binary regression analysis and in survival analysis. *International Statistical Review*, 58:243–252, 1990.

W. Feller. *An Introduction to Probability Theory and its Applications.* John Wiley and Sons, New York, 1971.

P. Hougaard. Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, 71(1):75–83, 1984.

J.P. Klein. Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, 48:795–806, 1992.

S.A Murphy, A.J. Rossini, and A.W. van der Vaart. Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, 92(439):968–976, 1997.

S.A. Murphy and A.W. van der Vaart. Semiparametric likelihood ratio inference. *The Annals of Statistics*, 25:1471–1509, 1997.

S.A. Murphy. Consistency in a proportional hazards model incorporating a random effect. *The Annals of Statistics*, 22(2):712–731, 1994.

S.A. Murphy. Asymptotic theory for the frailty model. *The Annals of Statistics*, 23(1):182–198, 1995.

S.A. Murphy. On profile likelihood. *Journal of the American Statistical Association*, 95:449–485, 2000.

G.G. Nielsen, R.D. Gill, P.K. Andersen, and T.I. Sorensen. A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics*, 19:25–43, 1992.

W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. *Numerical Recipies in Pascal. The Art of Scientific Computing.* Cambridge University Press, New York, NY, 1994.

A. Tsodikov. Semiparametric models: a generalized self-consistency approach (in press). *Journal of the Royal Statistical Society, Series B*, 2003.

A.W. van der Vaart. *Asymptotic statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, UK, 1998.

J.W. Vaupel, K.G. Manton, and E. Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16:439–454, 1979.

J.T. Wassel and M.L. Moeschberger. A bivariate survival model with modified gamma frailty for assessing the impact of interventions. *Statistics in Medicine*, 12:241–248, 1993.

12 Appendix

List of papers presented in the appendix

1. Tsodikov, A. (2003) Semiparametric models: A generalized self-consistency approach, *Journal of the Royal Statistical Society*, Series B, Vol. 65, 759-774.

2. Tsodikov, A., Ibrahim, J.G., and Yakovlev, A.Y. (2003) Estimating Cure Rates from Survival Data: An Alternative to Two-Component Mixture Models, *Journal of the American Statistical* Association, (review paper) to appear in December 2003.

3. Boucher, K., Asselain, B., Tsodikov, A., Yakovlev, A. Semiparametric versus parametric regression analysis based on the Bounded Cumulative Hazard Model: An application to breast cancer recurrence, In Semiparametric Models in Survival Analysis, Quality of Life and Reliability, Birkhauser (invited paper), to appear.

J. R. Statist. Soc. B (2003)
65, *Part* 3, *pp.* 759–774

Semiparametric models: a generalized self-consistency approach

A. Tsodikov

University of Utah, Salt Lake City, USA

[Received December 2002. Revised February 2003]

Summary. In semiparametric models, the dimension d of the maximum likelihood problem is potentially unlimited. Conventional estimation methods generally behave like $O(d^3)$. A new $O(d)$ estimation procedure is proposed for a large class of semiparametric models. Potentially unlimited dimension is handled in a numerically efficient way through a Nelson–Aalen-like estimator. Discussion of the new method is put in the context of recently developed minorization–maximization algorithms based on surrogate objective functions. The procedure for semiparametric models is used to demonstrate three methods to construct a surrogate objective function: using the difference of two concave functions, the EM way and the new quasi-EM (QEM) approach. The QEM approach is based on a generalization of the EM-like construction of the surrogate objective function so it does not depend on the missing data representation of the model. Like the EM algorithm, the QEM method has a dual interpretation, a result of merging the idea of surrogate maximization with the idea of imputation and self-consistency. The new approach is compared with other possible approaches by using simulations and analysis of real data. The proportional odds model is used as an example throughout the paper.

Keywords: EM algorithm; Frailty; Nonparametric maximum likelihood estimation; Profile likelihood; Semiparametric models

1. Introduction

Potentially unlimited dimension has been the most critical deterrent to the use of maximum likelihood estimation (MLE) in semiparametric regression models. In survival analysis, methods based on the partial likelihood (Cox, 1972) are specific to the proportional hazards (PH) model and do not extend to other models. Straightforward Newton-type methods of maximizing the likelihood for the full model generally require $O(d^3)$ operations to solve the system of score equations, where d is the number of model parameters. The principal part of the set of d parameters in a semiparametric model is used to specify a stepwise function H which approaches a continuous 'true' H in probability, as $d \to \infty$. Although theoretically almost any likelihood can be maximized by a Newton-type method, its high complexity makes the problem computationally difficult for large d. The development of general, stable and numerically efficient algorithms for semiparametric MLE has been a long-standing problem (Fleming and Lin, 2000). Such algorithms are the subject of this paper. The argument goes as follows. The bottle-neck of a maximization algorithm for a semiparametric likelihood is the estimation of H. Let l be the log-likelihood of a semiparametric model, treated as a functional of H. Consider a class of continuous semiparametric models with the log-likelihood of the form (informally)

$$l = \sum_t D_t \log\{\mathrm{d}H(t)\} + \sum_t \log\{\vartheta(H, t|z)\}, \tag{1}$$

Address for correspondence: A. Tsodikov, Division of Biostatistics, Huntsman Cancer Institute, University of Utah, 2000 Circle of Hope, Salt Lake City, UT 84112-5550, USA.
E-mail: atsodiko@hci.utah.edu

1369–7412/03/65759

where D_t is the number of exact observations at t (failures), z is a vector of covariates and $\vartheta > 0$ is some functional of H. The basic assumption that contributes to equation (1) is that the probability of failure in $[t, t + \mathrm{d}t]$ is proportional to $\mathrm{d}H(t)$, which is differentiability. To obtain an estimator for H, we differentiate l with respect to the set of $\{\mathrm{d}H(\tau)\}$. Informally, we arrive at the so-called self-consistency equation

$$\mathrm{d}H(\tau) = D_\tau \Big/ \sum_t \Theta(H, t|z), \tag{2}$$

where Θ is a functional representing a negative 'derivative' of $\log(\vartheta)$. Since both sides of equation (2) depend on H, an iterative procedure is required to make the equation self-consistent,

$$\mathrm{d}H^{(k+1)}(\tau) = D_\tau \Big/ \sum_t \Theta(H^{(k)}, t|z), \tag{3}$$

where k counts iterations. Iterative updating of H by using equation (3) is the basic idea behind the algorithm. As we shall see, the above procedure is intimately linked to the EM algorithm as used to fit certain PH frailty models in survival analysis (Oakes, 1989; Klein, 1992; Nielsen *et al.*, 1992). The EM algorithm handles H in an $O(d)$ way through the use of the Nelson–Aalen–Breslow estimator (Andersen *et al.*, 1993) for the cumulative hazard H. This is made possible as the M-step reduces to the PH model. However, a large amount of analytic work would be required to specify an estimation procedure for a new non-PH model. Expectation at the E-step may prove to be inaccessible in a closed form, and Monte Carlo extensions of the EM approach are much less computationally attractive. Recently, an optimization transfer approach (Lange *et al.*, 2000) was proposed that allows us to construct EM-like procedures without the use of missing data. For a target function $l(\mathbf{x}), \mathbf{x} \in \mathbb{R}^n$, the minorization–maximization (MM) algorithm (Lange *et al.*, 2000) proceeds by construction of the so-called surrogate objective function $Q(\mathbf{x}|\mathbf{y})$ such that $Q(\mathbf{y}|\mathbf{y}) = l(\mathbf{y})$, and $Q(\mathbf{x}|\mathbf{y}) \leqslant l(\mathbf{y})$, for any $\mathbf{x}$, to ensure monotonicity of the procedure. Maximization of the target function l proceeds iteratively as

$$\mathbf{x}^{(k+1)} = \arg\max_{\mathbf{x}} \{Q(\mathbf{x}|\mathbf{x}^{(k)})\}. \tag{4}$$

The MM algorithm converges in l and in $\mathbf{x}$ under fairly general conditions (Lange *et al.*, 2000). In the likelihood interpretation, the EM algorithm is a particular case of the MM algorithm. Unfortunately, there is no automatic way to construct Q. The procedure (3) interpreted as an MM algorithm is used to highlight three methods to construct a surrogate objective function: using the difference of two concave functions, the EM way and a new quasi-EM (QEM) approach. These methods link the EM algorithm for frailty models and its modifications with the MM algorithms. In the QEM approach, 'E' in the EM is replaced by the quasi-expectation operator QE, which is not based on the concept of a random variable. The result is the so-called QEM algorithm, which presents us with a recipe of generalizing an EM procedure into a 'distribution-free' one, representing a particular MM algorithm.

2. Profile likelihood approach

The problem of nonparametric maximum likelihood estimation (NPMLE) with the semiparametric model is to find estimates of regression coefficients β and an NPMLE estimate of H such that they deliver the maximum of a suitably defined likelihood function $l = l(\beta, H)$. In this paper we use a profile likelihood approach to maximize l. The profile likelihood is defined as a supremum of the full likelihood taken over the nonparametric part of the model

$$l_{\mathrm{pr}}(\beta) = \max_H \{l(\beta, H)\}. \tag{5}$$

Assuming that we can find the global maximum of l with respect to H, given β, we may write the profile likelihood as an implicit function of β

$$l_{\mathrm{pr}}(\beta) = l\{\beta, H(\beta)\}, \tag{6}$$

where $H(\beta)$ is the solution of equation (5). Our algorithms will be designed following a straightforward nested procedure:

(a) maximize $l_{\mathrm{pr}}(\beta)$ by a conventional non-linear programming method (e.g. a directions set method);
(b) for any β as demanded in the above maximization procedure, solve problem (5).

As the number of parameters of a semiparametric model is potentially unlimited, obtaining the inverse of the full information matrix can be computationally prohibitive. Therefore, we use the profile information matrix

$$\mathbf{I}^{\mathrm{P}}_{\beta,\beta} = -\frac{\partial^2 l_{\mathrm{pr}}\{\beta l_{\mathrm{pr}}(\beta)\}}{\partial\beta\,\partial\beta^{\mathrm{T}}} \tag{7}$$

to derive a standard error estimator for β. In this paper we adopt a pragmatic numerical approach. In the course of maximization of the profile likelihood with respect to β, a dense sample of the profile likelihood surface is generated near a stationary point. The curvature of the profile likelihood surface at the stationary point can be estimated by fitting a quadratic function to some domain around the point by using least squares. For example, the domain can be limited to points that cannot be rejected by using the likelihood ratio test (applied informally). Alternatively, a more sophisticated approach can be used based on implicit differentiation of l_{pr}.

The rest of the paper will be devoted to constructing efficient NPMLE methods for obtaining l_{pr}, i.e. for maximizing l with respect to H, given β, as this is the crux of the matter.

Practically, inference based on the profile likelihood is similar to that based on the partial likelihood for the PH model, which is quite convenient. Theoretically, however, inference based on the profile likelihood is not straightforward, as the usual theory of MLE does not apply to unlimited dimension. Important results have been obtained regarding a theoretical justification for the NPMLE method and the profile likelihood for semiparametric models (Murphy, 2000; van der Vaart, 1998; Murphy and van der Vaart, 1997). It was shown that profile likelihoods with nuisance parameters estimated out behave like ordinary likelihoods under some conditions. In particular, these results apply to the PH model, the proportional odds (PO) model (Murphy, 2000; Murphy *et al.*, 1997) and the PH frailty model (Murphy, 1994, 1995), and presumably to most other models.

Let $t_i, i=1,\ldots,n$, be a set of times, arranged in increasing order, and define $t_{n+1} := \infty$. Associated with each t_i is a set of individuals $\mathcal{D}_i$ with time-independent covariates $\mathbf{z}_{ij}, j \in \mathcal{D}_i$, who fail at t_i, and a similar set of individuals $\mathcal{C}_i$ with covariates $\mathbf{z}_{ij}, j \in \mathcal{C}_i$, who are censored at t_i. The observed event $\mathcal{E}_{ij}$ for the subject ij is a triple $(t_i, \mathbf{z}_{ij}, c_{ij})$, where c is a censoring indicator: $c=1$ if failure; $c=0$ if right censored. For any function $A(t)$, let $A_i = A(t_i)$, $\Delta A_i = |A(t_i) - A(t_i - 0)|$. A stepwise function H can be characterized by two vectors $\Delta\mathbf{H} = (\Delta H_1, \ldots, \Delta H_n)^{\mathrm{T}}$ and $\mathbf{t} = (t_1, \ldots, t_n)^{\mathrm{T}}$. With this notation, the likelihood of survival data under non-informative censoring takes the form

$$l = \sum_{i=1}^{n} D_i \log(\Delta H_i) + \sum_{i=1}^{n} \sum_{j \in \mathcal{C}_i \cup \mathcal{D}_i} \log\{\vartheta(\Delta\mathbf{H}, t_i | \beta, \mathbf{z}_{ij}, c_{ij})\}, \tag{8}$$

where D_i is the number of failures that are associated with t_i, and the function ϑ will be specified later for the class of non-linear transformation models (NTMs).

3. EM algorithm for a semiparametric model

For example, consider a PO model for the survival function G, given covariates $\mathbf{z}$,

$$G(t|\beta, \mathbf{z}) = G\{t|\theta(\beta, \mathbf{z})\} = \frac{\theta(\beta, \mathbf{z})}{\theta(\beta, \mathbf{z}) + H(t)}, \tag{9}$$

where θ is a predictor and H is some nonparametrically specified base-line cumulative hazard. The model is named after the PO property that for any two values of the predictor, θ_1 and θ_2, with corresponding survival functions $G_i(t) = G(t|\theta_i), i = 1, 2$, the odds ratio

$$\frac{\text{odds}\{G_1(t)\}}{\text{odds}\{G_2(t)\}} = \frac{\theta_1}{\theta_2}$$

is a constant in t, where $\text{odds}(a) = a/(1 - a)$.

This paper was inspired by the idea of representing a semiparametric model as a mixture (frailty) model, and to use the EM algorithm to fit it. With this idea in mind, consider a PH mixture model

$$G(t|\beta, \mathbf{z}) = E\{F(t)^{U(\beta, \mathbf{z})}|\mathbf{z}\}, \tag{10}$$

where $F = \exp(-H)$ is the base-line survival function corresponding to H, and $U = U(\beta, \mathbf{z})$ is used to indicate that the distribution of random variable U depends on covariates and regression coefficients. This model can be considered a compact expression for a family of so-called PH frailty models, or PH models with random effects considered by Hougaard (1984), Klein (1992), Nielsen *et al.* (1992), Wassel and Moeschberger (1993), Clayton and Cuzick (1985) and many others, for different distributions of U, possibly dependent on covariates.

To construct the EM algorithm for a particular model (PO in the example), we represent it as a PH mixture model (inverse transform), and then follow the usual logic of the EM algorithm construction for frailty models, as for example in Nielsen *et al.* (1992).

3.1. Inverse transform

We note that $\mathcal{L}(s|\cdot) = E[\exp\{-s\,U(\cdot)\}]$ is the Laplace transform of $U(\cdot)$, and that for the PH mixture model (10)

$$G(t|\cdot) = \mathcal{L}\{H(t)|\cdot\} = \mathcal{L}[-\log\{F(t)\}|\cdot].$$

From the latter equation and equation (9), we conclude that U for the PO model represents exponential regression, as $\mathcal{L}(s|\cdot) = \theta(\cdot)/\{\theta(\cdot) + s\}$ is the Laplace transform of an exponential distribution with mean θ^{-1}.

3.2. Complete-data likelihood

With the PH mixture model (10), pretend that U is known for each subject ij, continuing the notation of Section 2. The complete-data likelihood under non-informative right censoring corresponds to the PH model with predictors U_{ij}

$$l_{\text{cd}} = \sum_{i=1}^{n} \left\{ D_i \log(\Delta H_i) - \sum_{j \in \mathcal{C}_i \cup \mathcal{D}_i} U_{ij} H_i \right\}. \tag{11}$$

3.3. E-step

Since the complete-data likelihood (11) is linear in missing data U_{ij}, the E-step reduces to imputation of each U by the corresponding $\hat{U}$, the conditional expectation of U, given the observed event. Using the exponential distribution of U with mean θ^{-1}, after a little algebra, we obtain

$$\hat{U} = \frac{\int u\, F^{u}(uh)^{c}\theta\, \exp(-\theta u)\,\mathrm{d}u}{\int F^{u}(uh)^{c}\theta \exp(-\theta u)\,\mathrm{d}u} = \frac{\Gamma(c+2)\theta/(\theta+H)^{c+2}}{\Gamma(c+1)\theta/(\theta+H)^{c+1}} = \frac{c+1}{\theta(\beta,\mathbf{z})+H(t)}, \qquad (12)$$

where h is the hazard function corresponding to H. A similar derivation of $\hat{U}$ for a gamma frailty model can be found, for example, in Parner (1998).

3.4. M-step

Maximization of the complete-data likelihood (11) with respect to H, and with U_{ij} imputed by $\hat{U}_{ij}$, results in the Nelson–Aalen estimator

$$\Delta H_m = D_m \Big/ \sum_{ij \in \mathcal{R}_m} \hat{U}_{ij}, \qquad m = 1,\ldots,n,$$

where $\mathcal{R}_m = \{ij : j \in \mathcal{D}_i \cup \mathcal{C}_i, i \geqslant m\}$ is the set of subjects at risk just before t_m.

3.5. EM procedure for the proportional odds model

Finally, for the PO model we have the iterative EM procedure

$$\Delta H_m^{(k+1)} = D_m \left\{ \sum_{ij \in \mathcal{R}_m} \frac{c_{ij}+1}{\theta(\beta,\mathbf{z}_{ij}) + H_i^{(k)}} \right\}^{-1}, \qquad m = 1,\ldots,n, \qquad (13)$$

where k counts iterations.

3.6. Alternative derivation of procedure (13)

It is intriguing that we can formally derive procedure (13) as an immediate corollary of the argument presented in Section 1. Indeed, using equation (9), we write the likelihood for the PO model as

$$l = \sum_{i=1}^{n} D_i \,\log(\Delta H_i) + \sum_{j \in \mathcal{C}_i \cup \mathcal{D}_i} \log\left[\frac{\theta(\beta,\mathbf{z}_{ij})}{\{\theta(\beta,\mathbf{z}_{ij}) + H_i\}^{c_{ij}+1}} \right], \qquad (14)$$

On differentiating equation (14) with respect to ΔH_m, and assigning the iteration index k as in equations (1)–(3), we obtain expression (13).

This observation deserves discussion. The EM derivation presented above for the PO model is model specific and its feasibility depends on the success and simplicity of the inverse Laplace transform and the integrals that are evaluated at the E-step (12). The PH mixture representation of a semiparametric model may not exist, in which case the EM derivation ultimately fails. Necessary and sufficient conditions for this representation to exist are a corollary of the Bernstein theorem (Feller, 1971): the survival function must be a completely monotonic function of H. A function $\psi(H)$ is called completely monotonic if all its derivatives $\psi^{(i)}$ exist, $i=1,2,\ldots,$ and $(-1)^{i}\,\psi^{(i)}(H) \geqslant 0$, $H>0$. In particular, the survival function (10) of the PH mixture model is an infinitely differentiable function of F. The alternative derivation of procedure (13) bypasses all the above-mentioned difficulty and formally works for any model in a straightforward and simple fashion. This raises a series of questions. Does the procedure of Section 1 work for any model? What is its relationship to the EM algorithm? Does it inherit the monotonicity, stability and convergence of the EM algorithm?

A clue to generalizing the EM algorithm described above is the observation that the derivation of the E-step (12) does not require knowledge of the distribution of U. Indeed, denote by $\gamma(x)$ the moment-generating function of U (other arguments are omitted), so that $\gamma(x)=E(x^U)=\mathcal{L}\{-\log(x)\}$. Observe that the first equation in expression (12) can be written as

$$\hat{U}=\frac{E(F^U U^{c+1})}{E(F^U U^c)}=c+F\frac{\gamma^{(c+1)}(F)}{\gamma^{(c)}(F)}, \qquad (15)$$

where $\gamma^{(c)}$ denotes the derivative of order c; $\gamma^{(0)}:=\gamma$, $c=0,1$. Expression (15) represents a variation on the topic of the derivation of moments from the transform of a distribution. The consequence of equation (15) is a straightforward and general specification of the E-step for any mixture model formulated in terms of the moment-generating function. In fact, it is even more general as will be shown in what follows. To elaborate further on the issues raised above, we need to make the few theoretical observations considered in the next section.

4. General concepts

4.1. Construction using the difference of two concave functions

For studying procedure (3), the following MM construction (Lange *et al.*, 2000) is useful. Let $l(\mathbf{x})=B(\mathbf{x})-A(\mathbf{x})$, where A and B are differentiable concave functions. The iterative maximization procedure,

$$\nabla B(\mathbf{x}^{(k+1)})=\nabla A(\mathbf{x}^{(k)}), \qquad (16)$$

where $\nabla A(\mathbf{x})=\partial A/\partial\mathbf{x}$ is the gradient of A, represents an MM algorithm, as follows from convexity arguments. The surrogate objective function for the above construction has the form

$$Q(\mathbf{x}|\mathbf{x}^{(k)})=B(\mathbf{x}^{(k)})-A(\mathbf{x})+\nabla^{\mathrm{T}}A(\mathbf{x}^{(k)})(\mathbf{x}-\mathbf{x}^{(k)}). \qquad (17)$$

4.2. EM construction

Let $\mathcal{E}$ be the observed event, and U be a random variable (vector) representing missing data. The EM algorithm is a method to maximize the log-likelihood function $l(\mathbf{x})=\log\{L(\mathbf{x})\}$ of the form

$$L(\mathbf{x})=E\{L_0(\mathbf{x}|U)\}, \qquad (18)$$

where $L_0(\mathbf{x}|U)$ is the conditional likelihood, given missing data (the likelihood constructed to estimate $\mathbf{x}$ as if U were a covariate), and $\mathbf{x}$ does not include parameters of the distribution of U. To facilitate 'distribution-free' generalizations, we intentionally avoid explicit expressions involving the distribution of U, and we use the conditional rather than the joint likelihood of U and $\mathcal{E}$ to represent the EM procedure. In this construction, unknown parameters entering the distribution of U do not participate in the procedure, and the maximization is considered with respect to parameters $\mathbf{x}$ only. For any function $f(U)$, the conditional expectation of $f(U)$, given observed event $\mathcal{E}$ and $\mathbf{x}$, is represented as

$$E\{f(U)|\mathcal{E},\mathbf{x}\} = \frac{E\{f(U)\,L_0(\mathbf{x}|U)\}}{E\{L_0(\mathbf{x}|U)\}}.$$

This suggests the following explicit functional notation for the conditional expectation operator

$$E(f|g) := \frac{E(fg)}{E(g)}, \tag{19}$$

for any functions f and g of U, where U is a random variable, and $E(g)$ is the probability of the condition. A standard Jensen inequality argument shows that, with this notation,

$$Q(\mathbf{x}|\mathbf{y}) = l(\mathbf{y}) + E\left\{l_0(\mathbf{x}|U) - l_0(\mathbf{y}|U)|L_0(\mathbf{y}|U)\right\}, \qquad l_0 = \log(L_0), \tag{20}$$

is a surrogate objective function for the target function $l(\mathbf{x})$. The operation of finding $\hat{U}$ such that $l_0(\mathbf{x}|\hat{U}) = E\{l_0(\mathbf{x}|U)|L_0(\mathbf{y}|U)\}$ is referred to as missing data imputation. If imputation is easy (E-step), maximization of Q with respect to $\mathbf{x}$ reduces to that of $l_0(\mathbf{x}|\hat{U})$, a complete-data problem.

To prove that any converging sequence $\mathbf{x}^{(k)} \to \mathbf{x}^*$, designed according to equation (4), gives us a stationary point in the limit, we follow the argument at $\mathbf{x} = \mathbf{y} = \mathbf{x}^*$

$$\frac{\partial Q(\mathbf{x}|\mathbf{y})}{\partial \mathbf{x}} = \frac{E\{\partial L_0(\mathbf{x}|U)/\partial \mathbf{x}\}}{L(\mathbf{y})} = \frac{1}{L(\mathbf{y})}\frac{\partial L(\mathbf{x})}{\partial \mathbf{x}} = 0, \tag{21}$$

which implies that the score equation $\partial L(\mathbf{x})/\partial \mathbf{x} = 0$ is satisfied in the limit.

The EM algorithm proceeds by iterations $E\{l_0'(\mathbf{x}^{(k+1)}|U)\,L_0(\mathbf{x}^{(k)}|U)\} = 0$, where at the kth iteration this equation is solved for $x^{(k+1)}$.

4.3. Quasi-EM construction

Let us revisit the EM construction under the question what properties of the E-operator did we actually use in the derivation of Section 4.2? They are conditional expectation performed according to expression (19), linearity and the Jensen inequality in equation (20) and interchangeability of E and $\partial/\partial \mathbf{x}$ in equation (21). Operators satisfying these properties will be called QE. As soon as a QE operator satisfying these requirements and such that $L(\mathbf{x}) = \mathrm{QE}\{L_0(\mathbf{x}|U)\}$ is constructed, the likelihood function L can be maximized by an MM algorithm with the surrogate objective function as in equation (20) with E replaced by QE. The rationale behind this substitution is that the evaluation of E requires that U be a random variable, and that we know its distribution, whereas that of QE does not.

Formally, let $\mathcal{B}$ be some set of basis functions (including the function $f(u) \equiv 1$), and $\mathcal{S}$ be a set of all admissible functions stretched on $\mathcal{B}$ using linear combinations. In other words $\mathcal{S}$ consists of functions f such that $f = \Sigma_i a_i f_i$ for any sequence (finite or infinite) of functions $\{f_i\}$, $f_i \in \mathcal{B}$, and real numbers $\{a_i\}$.

Define QE as a linear functional mapping $\mathcal{S}$ to real numbers such that

(a) $\mathrm{QE}(\mathbf{1}) := 1$, where $\mathbf{1}$ means a function that is equivalent to 1 (*normalization*),
(b) for any $f = \Sigma_i a_i f_i \in \mathcal{S}$, $f_i \in \mathcal{B}$, $\mathrm{QE}(f) := \Sigma_i a_i \mathrm{QE}(f_i)$ (*linearity*),
(c) for any function $f(u,a) \in \mathcal{S}$, and such that $\partial f(u,a)/\partial a \in \mathcal{S}$,

$$\frac{\partial \mathrm{QE}\{f(U,a)\}}{\partial a} = \mathrm{QE}\left\{\frac{\partial f(U,a)}{\partial a}\right\}$$

(*interchangeability*),

(d) as in expression (19), for any functions g and $fg \in \mathcal{S}$ and such that $\mathrm{QE}(g) \neq 0$,

$$\mathrm{QE}(f|g) := \frac{\mathrm{QE}(fg)}{\mathrm{QE}(g)} \tag{22}$$

(*conditional QE*) and

(e) given the functions g, fg and $g\log(f) \in \mathcal{S}$,

$$\mathrm{QE}\{\log(f)|g\} \leqslant \log\{\mathrm{QE}(f|g)\} \tag{23}$$

(*Jensen inequality*).

Let us consider the QE requirements more closely. We start by postulating one basis function $f_0(U,a)$ and the value of QE on that basis function $\mathrm{QE}(f_0) := \gamma(a)$, where γ is some function of a. Dependent on how many times we are allowed to differentiate under the QE symbol, the derivatives

$$f^{(i)}(U,a) = \partial^{(i)} f_0(U,a)/\partial a^i$$

must also be included in the set of admissible functions $\mathcal{S}$, so that derivatives of $\mathrm{QE}(f_0)$ can be defined. Moreover, QE on $f^{(i)}$, $i = 1,2,\ldots$, are automatically defined by the interchangeability property through derivatives of γ, as $\mathrm{QE}\{f^{(i)}\} = \gamma^{(i)}$. As we can see, QE construction is cloned from f_0 and γ. Mathematical expectation $E\{f_0(U,a)\}$ is an integral transform of U, where a is an argument of the transform. Dependent on the choice of the function f_0, QE mimics differential properties of the corresponding transform, whereas the function γ is not necessarily an integral transform.

To study procedure (3), a QE construction based on moment-generating functions is useful. This construction is cloned from the basis function $f_0(U,a) = a^U$, $0 \leqslant a \leqslant 1, U \geqslant 0$. For a mathematical expectation, $E(f_0) = \gamma$ is the moment-generating function of U. If we want to be able to differentiate twice under the QE symbol, we need two more basis functions Ua^U and U^2a^U, so that

$$\left.\begin{aligned} \mathrm{QE}(a^U) &= \gamma(a), \\ \mathrm{QE}(Ua^U) &= a\gamma'(a), \\ \mathrm{QE}(U^2a^U) &= a\gamma'(a) + a^2\gamma''(a), \end{aligned}\right\} \tag{24}$$

by the interchangeability property with two derivatives of γ allowed. We now derive the Jensen inequality for this construction. First, noting equation (15), introduce the conditional QE

$$\Theta(x|c) := \mathrm{QE}(U|U^c x^U) = c + x\frac{\gamma^{(c+1)}(x)}{\gamma^{(c)}(x)}, \tag{25}$$

where we used expressions (22) and (24) to obtain the right-hand part of expression (25), $c = 0, 1$. The function Θ is a surrogate of conditional expectation of U, given observed data, and it serves as an imputation operator in the QEM construction. Next, let $\mathcal{B}_k$ be a family of functions $\mathcal{B}_k = \{U^k a^U, 0 \leqslant a \leqslant 1, U \geqslant 0\}$. Then by using the fact that, for $f = f(U,a) = a^U$ and $g = g(U,b) = U^c b^U$,

$$\frac{\partial}{\partial a}[\mathrm{QE}\{\log(f)|g\} - \log\{\mathrm{QE}(f|g)\}] = \frac{1}{a}\{\Theta(b|c) - \Theta(ab|c)\},$$

the following can be proved.

Theorem 1 (Jensen inequality for QE). Let $\Theta(x|c)$ as defined by expression (25) be a non-decreasing function of x, $c=0,1$. Then inequality (23) holds true for any $f \in \mathcal{B}_0$ and $g \in \mathcal{B}_0 \cup \mathcal{B}_1$.

It is interesting that mathematical expectation satisfies the assumption of non-decreasing Θ, which we call the *convexity assumption* for reasons that will become clear later as we discuss an application of the above theory to semiparametric likelihood.

Theorem 2 (convexity for mathematical expectation). Let γ be a function defined by using the mathematical expectation operator E as $\gamma(x) := E(x^U)$, where U is a non-negative random variable. Then $\Theta(x|c)$ as defined in expression (25) is non-decreasing in x for any $c = 0, 1$.

Proof. The Cauchy–Schwartz inequality with functions $\xi(U,x) = U^{1+c/2}x^{U/2}$ and $\zeta(U,x) = U^{c/2}x^{U/2}$ can be used to show that $\Theta'(x|c) \geqslant 0$.

5. Non-linear transformation models

5.1. Definition

To study procedure (1)–(3) in more detail using the developments of Section 4, we need to specify a certain structure of the likelihood function to be optimized. To do this, let us confine ourselves to a large, yet specific, class of semiparametric survival models. Consider a parametric regression model with support on [0, 1]. Let $\gamma(x|\boldsymbol{\beta},\mathbf{z}), x \in [0,1]$, be a parametrically specified distribution function in x, conditional on covariates $\mathbf{z}$. We require that γ be twice differentiable with respect to x and regression coefficients $\boldsymbol{\beta}$.

We can now define a semiparametrically specified survival function $G(t|\boldsymbol{\beta},\mathbf{z})$, given covariates, as

$$G(t|\boldsymbol{\beta},\mathbf{z}) = \gamma\{F(t)|\boldsymbol{\beta},\mathbf{z}\}, \tag{26}$$

where the base-line survival function F is specified nonparametrically. The class of models (26) will be called NTMs, to give it a name. Functions like γ will be called NTM-generating functions. An NTM is obtained by plugging a nonparametrically specified survival function F into a parametric distribution function γ with the support compatible with the range of F. One important subclass of NTMs is the family of PH mixture models (10), for which γ is a moment-generating function of U,

$$\gamma(x|\boldsymbol{\beta},\mathbf{z}) = E(x^{U(\boldsymbol{\beta},\mathbf{z})}|\mathbf{z}).$$

To represent a semiparametric model in the NTM form, we need to express its survival function G as a function γ of a base-line survival function F (this representation is not unique and is not always possible). For example, from equation (9) we obtain the PO model in the form $G(t|\cdot) = \theta(\cdot)/[\theta(\cdot) - \log\{F(t)\}]$, which gives

$$\gamma(x|\cdot) = \frac{\theta(\cdot)}{\theta(\cdot) - \log(x)}. \tag{27}$$

The class of NTMs includes the class of linear transformation models (Cheng *et al.*, 1995, 1997). It is easy to show that a linear transformation model can be represented as $\gamma(x|\boldsymbol{\beta},\mathbf{z}) = p[\log\{\theta(\boldsymbol{\beta},\mathbf{z})\} + q(x)]$, where p is a parametrically specified tail function, q is an inverse tail function and θ is a predictor (it is convenient to specify q as the inverse of p; then $\theta = 1$ corresponds to the base-line $\gamma(x|\cdot) = x$).

5.2. Algorithm

In the survival analysis formulation, under non-informative right censoring, the contribution of observations sampled from an NTM (26) to the likelihood are $\log(-\mathrm{d}G)$ and $\log(G)$, for a failure and censored observation respectively. We have

$$-\mathrm{d}G(t|\beta,\mathbf{z}) = \gamma'\{F(t)|\beta,\mathbf{z}\}\,F(t)\,\mathrm{d}H(t),$$

where $\gamma'(x|\cdot) = \partial\gamma(x|\cdot)/\partial x$, differentials are taken with respect to t under a continuous model and we recall that $F = \exp(-H)$. We may now rewrite the likelihood (8) for an NTM as

$$l = \sum_{i=1}^{n} D_i \log(\Delta H_i) + \sum_{i=1}^{n} \sum_{j\in\mathcal{C}_i\cup\mathcal{D}_i} \log\{\vartheta(F_i|\beta,\mathbf{z}_{ij},c_{ij})\}, \tag{28}$$

where

$$\vartheta(x|\cdot,c) = x^c\,\gamma^{(c)}(x|\cdot),$$

and ΔH_i is substituted for $\mathrm{d}H(t_i)$. It is easy to check that a negative derivative of $\log\{\vartheta(F_i|\cdot,c)\}$ with respect to ΔH_m is represented by $\Theta(F_i|\cdot,c)$, if $m \leqslant i$, and is equal to 0 otherwise, where the function Θ is defined by expression (25). Therefore, the construction (1)–(3) of Section 1 leads us to the iteration scheme

$$\Delta H_m^{(k+1)} = D_m \Big/ \sum_{ij\in\mathcal{R}_m} \Theta(F_i^{(k)}|\beta,\mathbf{z}_{ij},c_{ij}). \tag{29}$$

This procedure is a generalization of procedure (13) to the NTM family. For the PO model, substituting equation (27) into expression (25), we obtain

$$\Theta(x|\cdot,c) = \frac{c+1}{\theta(\cdot)-\log(x)}. \tag{30}$$

It is clear that, with Θ given by equation (30), the general procedure (29) turns into the procedure (13) derived for the PO model in Section 3.

5.3. Quasi-expectation form of a non-linear transformation model

We now make use of the QE theory of Section 4.3 to provide a link between NTMs and the QE operator. Equations (24) summarizing second-order differential properties of the QE operator will be the main instrument of this section.

First, let us synchronize the development of Section 4.3 and the definition of NTM (26) in Section 5.1 by assuming that the function γ that is used in both sections is the same function. In fact, we already used this synchronization when we noticed in the previous section that Θ in equation (29) and in expression (25) is the same function. Now, from the first line of expression (24), with $F(t)$ instead of a, we obtain the QE form of the NT model

$$G(t|\beta,\mathbf{z}) = \mathrm{QE}_{\beta,\mathbf{z}}\{F(t)^U\}, \tag{31}$$

where the subscript $\beta,\mathbf{z}$ to the QE operator indicates that QE is defined by using the function $\gamma(x|\beta,\mathbf{z})$. Equation (31) is a postulate in the definition of the QE operator, and its link to the NTMs is established as we assume that QE is defined by using an NTM-generating function γ.

Now, consider the likelihood function l (28). Given an observation $(t,\mathbf{z},c), c = 0, 1$, its contribution $v\{F(t)|\beta,\mathbf{z},c\}$ to the likelihood $L = \exp(l)$ can be written as

$$v(F|\cdot,c) = \vartheta(F|\cdot,c)\Delta H^c = \Delta H^c F^c\,\gamma^{(c)}(F|\cdot,c) = \mathrm{QE}(\Delta H^c U^c F^U),$$

where the last equation follows from the first two lines of expression (24) and linearity of QE. As a result, the likelihood of an NTM mimics that of a mixture model

$$L = \prod \mathrm{QE}(\Delta H^c U^c F^U).$$

Consider the hazard function $\lambda(t|\mathbf{z})$, corresponding to the survival function $G(t|\mathbf{z})$. Differentiating the survival function (26), and using expression (24), we have

$$\lambda(t|\cdot) = -\frac{1}{G(t|\cdot)}\frac{\partial G(t|\cdot)}{\partial t} = \frac{\gamma'\{F(t)|\cdot\}\,F(t)}{\gamma\{F(t)|\cdot\}}h(t) = \frac{\mathrm{QE}\{U\,F(t)^U\}}{\mathrm{QE}\{F(t)^U\}}h(t),$$

where h is the hazard function corresponding to F. Applying the definition of conditional QE (22) to this expression, and using expression (25), we obtain

$$\lambda(t|\mathbf{z}) = \mathrm{QE}\{U|F(t)^U\}\,h(t) = \Theta(F|\cdot,0)\,h(t).$$

This is a generalization of the fact that the population hazard function at time t in a heterogeneous population is represented as a conditional average, given survival up to time t.

Bringing these derivations together with expression (25), we have the following theorem.

Theorem 3 (QEM construction). Consider a survival analysis problem for an NTM generated by the function $\gamma(x|\beta,\mathbf{z})$, with fixed covariates. With the QE operator as defined in Section 4.3, and using the same NTM-generating function γ in its definition, the following representations are valid: *survival function*,

$$G(t|\beta,\mathbf{z}) = \mathrm{QE}_{\beta,\mathbf{z}}\{F(t)^U\};$$

hazard function,

$$\lambda(t|\mathbf{z}) = \mathrm{QE}\{U|F(t)^U\}\,h(t) = \Theta(F|\cdot,0)\,h(t),$$

where λ and h are hazards functions corresponding to G and F respectively; *likelihood function*,

$$l = \sum_{i=1}^{n}\left(\sum_{j\in\mathcal{C}_i\cup\mathcal{D}_i}\log[\mathrm{QE}_{\beta,\mathbf{z}_{ij}}\{(U\Delta H_i)^{c_{ij}}F_i^U\}]\right);$$

imputation operator,

$$\hat{U} = \mathrm{QE}(U|U^cF^U) = \Theta(F|\cdot,c), \qquad c = 0,1,$$

where $\hat{U}$ denotes U, imputed by using the conditional QE operator.

6. Summary and justification of the procedure

Let us now go back to the EM algorithm of Section 3 and see how the results obtained since then allow us to streamline and justify our algorithm construction, using the PO model as an example. In summary, we now have the following procedure.

(a) Obtain the NTM-generating function, representing the model survival function as a function of F. For the PO model (9) $G(t|\cdot) = \theta(\cdot)[\theta(\cdot) - \log\{F(t)\}]^{-1}$, we have equation (27),

$$\gamma(x|\cdot) = \theta(\cdot)\{\theta(\cdot) - \log(x)\}^{-1}.$$

(b) Obtain the imputation operator (25)

$$\Theta(x|\cdot) = c + x\,\gamma^{(c+1)}(x|\cdot)\,\gamma^{(c)}(x|\cdot)^{-1}.$$

For the PO model, this results in equation (30),

$$\Theta(x|\cdot) = (c+1)\{\theta(\cdot) - \log(x)\}^{-1}.$$

Check that $\Theta(x|\cdot)$ is a non-decreasing function of x (see the justification below).

(c) Obtain the profile likelihood by iterations (29),

$$\Delta H_m^{(k+1)} = D_m \left\{ \sum_{ij\in\mathcal{R}_m} \Theta(F_i^{(k)}|\beta, \mathbf{z}_{ij}, c_{ij}) \right\}^{-1}.$$

(d) Maximize the profile likelihood with respect to β as in Section 2.

For the PH mixture model, QE is equivalent to E (compare equations (31) and (10)), which makes Θ the conditional expectation of U, given observed data (compare expressions (25) and (15)). In this case, the above procedure is an EM algorithm.

Justification of this procedure works through the proof of monotonicity (i.e. the likelihood is improved at each step) under the following assumption.

6.1. Convexity assumption

Consider an NTM with the NTM-generating function γ. Assume that

$$\Theta(x|\beta, \mathbf{z}, c) \text{ is a non-decreasing function of } x, \text{for any } \beta, \mathbf{z}, c. \qquad (32)$$

We have two ways to show monotonicity.

(a) Observe that, under assumption (32), the likelihood (28), as a function of the vector $\Delta\mathrm{H}$, represents a difference between two concave functions $\Sigma_i D_i \log(\Delta H_i)$ and $-\Sigma_{ij} \log\{\vartheta(F_i|\cdot)\}$. This follows from the fact that Θ is the negative derivative of $\log(\vartheta)$ with respect to H. Therefore, monotonicity follows from the results of Section 4.1.

(b) Observe that the likelihood is represented as a QE $L = \Pi\,\mathrm{QE}(\Delta H^c U^c F^U)$ (Section 5.3), and that under assumption (32) the QE operator satisfies the Jensen inequality (Section 4.3). Therefore, the EM proof of monotonicity works.

Convergence of the algorithm under monotonicity follows from the results of Lange *et al.* (2000) and Wu (1983) under fairly general conditions.

7. Real data example

As an example we use data from the National Cancer Institute's 'Surveillance epidemiology and end results' programme. Using the publicly available database for the programme, 11621 cases of primary prostate cancer diagnosed in the state of Utah between 1988 and 1999 were identified. The following selection criteria were applied to a total of 19819 Utah cases registered in the database: valid positive survival time, valid stage of the disease and age 18 years or more. Prostate cancer specific survival was analysed by the stage of the disease (localized or regional *versus* distant). For the definition of stages as well as for other details of the data we refer the reader to the documentation for the programme at `http://seer.cancer.gov/`.

The PH and the PO models with $\mathbf{z}$ representing two groups corresponding to the localized or regional stage (10765 cases) and distant stage (856 cases) respectively were fitted by using the profile MM algorithm. The log-odds-ratio was estimated as $\hat{\beta} = -3.251$ with 95% asymptotic confidence interval $(-3.416, -3.086)$. A likelihood ratio test showed that the difference between groups is highly significant ($p < 0.0001$). Observed (Kaplan–Meier) and expected model-based

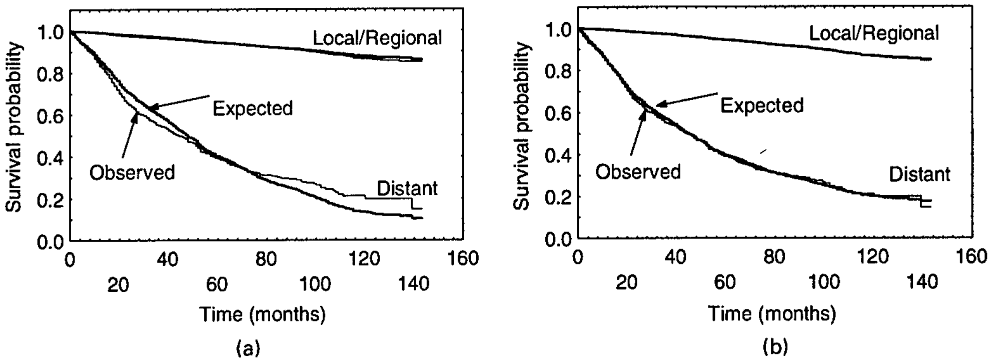

Fig. 1. Observed (———) *versus* expected (———) plots corresponding to (a) the PH and (b) the PO model fitted to the prostate cancer data

estimates of the survival functions by group are shown in Fig. 1. The PO model showed a superior fit to the data.

On the basis of the PO model, four different approaches to model fitting are compared.

(a) *MM or QEM*: the method is described in Section 6. Maximization of the profile likelihood is performed by the Powell method (Press *et al.*, 1994).

(b) *EM*: the EM method is similar to the EM algorithm as used to fit frailty models with predictor $\theta(\beta, \mathbf{z})$. Using the QEM formulation, the procedure is as follows.

   (i) With the current iteration $\beta^{(k)}$ and $F^{(k)}$ compute $V_{ij}^{(k)} = \theta^{-1}(\beta^{(k)}, \mathbf{z}_{ij})\,\Theta(F_i^{(k)}|\beta^{(k)}, \mathbf{z}_{ij})$ for each subject ij.

   (ii) Maximize the partial likelihood for a PH model with (imputed) predictor $V_{ij}^{(k)}\,\theta(\beta, \mathbf{z}_{ij})$ with respect to β. Set $\beta^{(k+1)}$ at the solution.

   (iii) Update the function F by using the Nelson–Aalen estimator for the PH model fitted at the previous step. Denote the solution by $F^{(k+1)}$.

   (iv) Set $k = k + 1$. Continue iterations until convergence.

(c) *Parametric*: in the parametric method, the function F is specified as a Weibull survival function. The parametric regression model is fitted by using the Powell method applied to all model parameters.

(d) *Full model (Powell)*: apply the Powell method to maximize the log-likelihood of the full semiparametric model with respect to the joint vector of regression coefficients β and the base-line hazard $\Delta\mathbf{H}$.

Computation of θ, Θ or γ is counted as one operation. For a given tolerance ε, the stopping rule is defined as $l^{(k+1)} - l^{(k)} < \varepsilon$. If the method required solving a nested numerical problem (MM or EM), the tolerance for the nested problem is specified as $\varepsilon/100$.

First, we evaluate the precision by operations characteristics of the above numerical methods. The precision is measured by $l^* - l^{(k)}$, where the exact solution l^* was approximated by the solution obtained for $\varepsilon = 10^{-20}$. Shown in Fig. 2 are the precision by operations curves for the four methods, obtained by varying ε. It is clear from Fig. 2 that the profile MM algorithm outperforms the other approaches in the number of operations that are required to reach a given precision. The profile MM method is closely followed by the frailty EM algorithm. Fitting the full semiparametric model by the Powell method shows the worst performance. The advantage of the EM-like approaches compared with methods that invoke the function F into a conventional maximization is explained by the utilization of a closed form solution for F in the

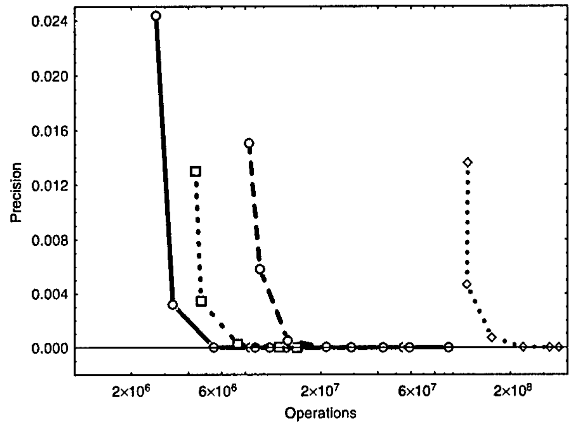

Fig. 2. Precision of likelihood maximization by the number of operations (precision is measured as the difference between the limiting value of the likelihood as operations tend to ∞ and the maximal likelihood value achieved under a stopping rule; curves closer to the *y*-axis correspond to more efficient numerical methods): ———, MM; -------, EM; – – –, parametric; ·······, full model

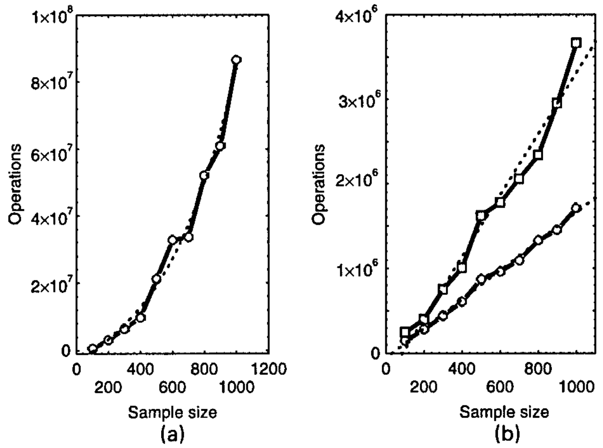

Fig. 3. Numerical efficiency by sample size for (a) the full model (·······, polynomial fit) and (b) the EM (□) and MM (○) methods (·······, linear fits): ———, number of operations needed to achieve a fixed precision by sample size (each point corresponds to a sample generated from a parametric PO model with a Weibull base-line survival function with parameters specified by using the model fit to data described in Section 7)

form of the Nelson–Aalen–Breslow estimator. For the same reason, the MM and the EM procedures show a linear trend with increasing dimension, given fixed precision as shown in Fig. 3. To obtain Fig. 3, samples of size 100–1000 were generated from the parametric (Weibull) PO model fitted to the same data. The MM, EM and full model (Powell) procedures were applied to each such sample. The tolerance $\varepsilon = 10^{-3}$ was used for the MM algorithm. The tolerance for the other two methods was tuned to give a likelihood that was as close as possible yet smaller than the likelihood achieved by the MM method (to keep the comparison conservative). The profile MM algorithm shows the most favourable behaviour with increasing dimension, followed by the EM procedure. It comes as no surprise that the full model (Powell) method shows the greatest complexity.

8. Conclusion

We presented an application of the general MM principle to a class of semiparametric models. Three methods of specifying the surrogate objective function were demonstrated. In particular, we clarified the connection between the likelihood-based MM principle and the imputation-based self-consistency principle that is used in EM algorithms for semiparametric models. To study this connection, we built an EM-like world behind the MM algorithm by using the QEM construction. The approaches were illustrated by using continuous NTMs in a survival analysis context. This is just one example of how these constructions can be used. Discrete survival models, cure models, multivariate semiparametric models, models with time-dependent covariates and many other statistical models can be treated by application of the principles presented in this paper. Construction of surrogate objective functions is not straightforward. Having an option to work a particular problem from both ends (likelihood or convexity *versus* imputation or self-consistency) may increase the chance of finding efficient and general procedures that are applicable to large classes of models.

Acknowledgements

The author thanks Dr Ken Boucher, Dr Marvin Zelen, the referees and the Joint Editor for many helpful comments. This research is partially supported by National Cancer Institute grant 1U01 CA97414-01 (Cancer Intervention and Surveillance Modeling Network), Department of Defence grant DAMD17-03-1-0034 and National Institute of Ageing grant 1RO1 AG14650-01A2.

References

Andersen, P., Borgan, Ø., Gill, R. and Keiding, N. (1993) *Statistical Models based on Counting Processes*. New York: Springer.

Cheng, S., Wei, L. and Ying, Z. (1995) Analysis of transformation models with censored data. *Biometrika*, **82**, 835–845.

Cheng, S., Wei, L. and Ying, Z. (1997) Predicting survival probabilities with semiparametric transformation models. *J. Am. Statist. Ass.*, **92**, 227–235.

Clayton, D. and Cuzick, J. (1985) Multivariate generalizations of the proportional hazards model (with discussion). *J. R. Statist. Soc.* A, **148**, 82–117.

Cox, D. (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B, **34**, 187–220.

Feller, W. (1971) *An Introduction to Probability Theory and Its Applications*. New York: Wiley.

Fleming, T. and Lin, D. (2000) Survival analysis in clinical trials: past developments and future directions. *Biometrics*, **56**, 971–983.

Hougaard, P. (1984) Life table methods for heterogeneous populations: distributions describing the heterogeneity. *Biometrika*, **71**, 75–83.

Klein, J. (1992) Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, **48**, 795–806.

Lange, K., Hunter, D. and Yang, I. (2000) Optimization transfer using surrogate objective functions (with discussion). *J. Comput. Graph. Statist.*, **9**, 1–59.

Murphy, S. (1994) Consistency in a proportional hazards model incorporating a random effect. *Ann. Statist.*, **22**, 712–731.

Murphy, S. (1995) Asymptotic theory for the frailty model. *Ann. Statist.*, **23**, 182–198.

Murphy, S. (2000) On profile likelihood. *J. Am. Statist. Ass.*, **95**, 449–485.

Murphy, S., Rossini, A. and van der Vaart, A. (1997) Maximum likelihood estimation in the proportional odds model. *J. Am. Statist. Ass.*, **92**, 968–976.

Murphy, S. and van der Vaart, A. (1997) Semiparametric likelihood ratio inference. *Ann. Statist.*, **25**, 1471–1509.

Nielsen, G., Gill, R., Andersen, P. and Sorensen, T. (1992) A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Statist.*, **19**, 25–43.

Oakes, D. (1989) Bivariate survival models induced by frailties. *J. Am. Statist. Ass.*, **84**, 487–493.

Parner, E. (1998) Asymptotic theory for the correlated gamma frailty model. *Ann. Statist.*, **26**, 183–214.

Press, W., Flannery, B., Teukolsky, S. and Vetterling, W. (1994) *Numerical Recipes in Pascal: the Art of Scientific Computing*. New York: Cambridge University Press.

van der Vaart, A. (1998) *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Wassel, J. and Moeschberger, M. (1993) A bivariate survival model with modified gamma frailty for assessing the impact of interventions. *Statist. Med.*, **12**, 241–248.

Wu, C. (1983) On the convergence properties of the EM algorithm. *Ann. Statist.*, **11**, 95–103.

Estimating Cure Rates From Survival Data: An Alternative to Two-Component Mixture Models

A. D. Tsodikov, J. G. Ibrahim, and A. Y. Yakovlev

This article considers the utility of the bounded cumulative hazard model in cure rate estimation, which is an appealing alternative to the widely used two-component mixture model. This approach has the following distinct advantages: (1) It allows for a natural way to extend the proportional hazards regression model, leading to a wide class of extended hazard regression models. (2) In some settings the model can be interpreted in terms of biologically meaningful parameters. (3) The model structure is particularly suitable for semiparametric and Bayesian methods of statistical inference. Notwithstanding the fact that the model has been around for less than a decade, a large body of theoretical results and applications has been reported to date. This review article is intended to give a big picture of these modeling techniques and associated statistical problems. These issues are discussed in the context of survival data in cancer.

KEY WORDS: Bayesian methods; Biologically based models; Bounded cumulative hazard; Cure models; Hazard regression; Semiparametric inference; Survival data.

1. INTRODUCTION

In many clinical and epidemiological settings, investigators observe cause-specific survival curves that tend to level off at a value strictly greater than 0 as time increases. A prominent example of this pattern is shown in Figure 1(a). A well-pronounced plateau in this display of the Kaplan–Meier curve may be thought of as an indication of the presence of a proportion of patients for whom the disease under study will never recur. Alternatively, one can consider such patients to be cured. Clearly, estimating the proportion of cured patients may have important medical implications. In addition, clinical covariates may exert dissimilar effects on the probability of cure and the timing of tumor relapse or other events of interest. This is apparent from Figure 1(b), where two survival curves for patients with localized breast cancer stratified by age are presented. These plots suggest that the two categories of patients are likely to have a similar probability of cure, whereas a short-term effect of age at diagnosis on cancer-specific survival is highly plausible. It is also clear that the commonly used proportional hazards model fails to fit the data shown in Figure 1(b). In Section 3.3 we present a detailed example involving prostate cancer survival that has major biomedical implications. Such an advance would have not been possible to make without invoking the concept of cure. The preceding examples suggest at least two advantages that survival models have for allowing for the presence of cured individuals: (1) They enrich our ability to interpret survival analysis in terms of characteristics that have a clear biomedical meaning; (2) they lead to more general regression models, thereby extending our ability to describe actual data. The latter contention holds whether the probability of cure is significantly separated from 0 (see Sec. 3).

To develop relevant methods of statistical data analysis, one has to provide a rigorous definition of cure rate. In this article, we will proceed from the most widely accepted concept of biological cure. The probability of (biological) cure, variously referred to as the cure rate or the surviving fraction, is defined as an asymptotic value of the survival function $\overline{G}(t)$ as t tends to ∞. Let X be the survival time with cumulative distribution function (c.d.f.) $G(t) = 1 - \overline{G}(t)$. Under a continuous model the existence of a nonzero surviving fraction, p, is determined by the behavior of the hazard function, $\lambda(t)$, by virtue of the equality

$$p = \lim_{t\to\infty} \overline{G}(t) = \exp\left\{-\int_0^\infty \lambda(u)\,du\right\}. \qquad (1.1)$$

Whenever $p > 0$ the underlying survival time distribution is said to be improper. Clearly, $\lambda(u) \to 0$ as $u \to \infty$ if $p > 0$ and the limit of $\lambda(u)$ (as $u \to \infty$) exists. Formulated in terms of the marginal failure time distribution, this definition does not imply that the overall survival time may be infinite, because the time to death from other causes (censoring time) is finite with probability 1. Therefore, the distribution of the observed lifetime in the presence of competing risks is always proper, and there is no defiance of common sense.

Boag (1949) and later Berkson and Gage (1952) proposed a two-component (binary) mixture model for the analysis of survival data when a proportion of patients are cured. Since then, the binary mixture-based approach has become the dominant one in the literature on cure models (Miller 1981; Farewell 1982; Goldman 1984; Greenhouse and Wolfe 1984; Gamel, McLean, and Rosenberg 1990; Gordon 1990; Bentzen, Johansen, Overgaard, and Thames 1991; Goldman and Hillman 1992; Kuk and Chen 1992; Laska and Meisner 1992; Maller and Zhou 1992, 1994, 1995, 1996; Sposto, Sather, and Baker 1992; Yamaguchi 1992; Gamel and Vogel 1993; Gamel, Vogel, Valagussa, and Bonadonna 1994; Chappell Nondahl, and Fowler 1995; Taylor 1995; Gamel, Meyer, Feuer, and Miller 1996; Peng and Dear 2000; Sy and Taylor 2000, 2001, to name a few). The main idea behind this approach is that any improper survival function can be represented as

$$\overline{G}(t) = \mathrm{E}\left\{[\overline{G}_0(t)]^M\right\} = p + (1-p)\overline{G}_0(t), \qquad (1.2)$$

A. D. Tsodikov is Associate Professor of Biostatistics, University of California, Davis, CA 95616 (E-mail: *atsodikov@ucdavis.edu*). J. G. Ibrahim is Professor, Department of Biostatistics, University of North Carolina, McGavran-Greenberg Hall, Chapel Hill, NC 27599 (E-mail: *ibrahim@bios.unc.edu*). A. Y. Yakovlev is Professor, Department of Statistics and Computational Biology, University of Rochester, 601 Elmwood Avenue, Box 630, Rochester, NY 14642 (E-mail: *andrei-yakovlev@urmc.rochester.edu*). The authors wish to thank the editor, the associate editor, and three referees for several suggestions and editorial changes which have greatly improved the paper. The research of A. Tsodikov and A. Yakovlev was supported by NIA/NIH grant 1 RO1 AG14650-01A2, NCI/NIH 1U01 CA88177-01, CA97414-01, and DAMD 17-03-1-0034. Dr. Ibrahim's research was partially supported by NCI/NIH CA 70101, CA 74015, and CA 23318.

Journal of the American Statistical Association
December 2003, Vol. 98, No. 464, Review Article
DOI 10.1198/016214503000001007

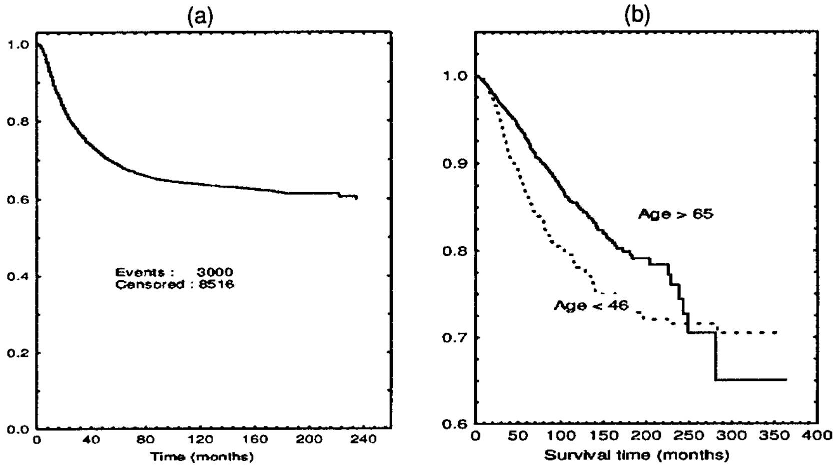

Figure 1. (a) Relapse-Free Survival for Patients With Hodgkin's Disease Treated by Radiotherapy. Data from the International Database on Hodgkin's Disease. (b) Breast cancer specific survival in local stage by age. Data from the Surveillance Epidemiology and End Results Program.

where M is a binary random variable taking on the values of 0 and 1 with probability p and $1 - p$, respectively, with

$$p = \Pr\{X = \infty\},$$

and $\overline{G}_0(t)$ is defined as the survival function for the time to failure conditional upon ultimate failure, that is,

$$\overline{G}_0(t) = \Pr\{X \geq t | X < \infty\}. \tag{1.3}$$

When designing regression counterparts of model (1.2), it is common practice to use the logistic regression model for incorporating covariates into the probability p, and proportional hazards regression for modeling the effect of the covariates on the conditional survival function $\overline{G}_0$.

An alternative, but equally general, representation of an improper survival time distribution can be obtained by assuming that the cumulative hazard $\Lambda(t) = \int_0^t \lambda(t)\,dt$ has a finite positive limit, say θ, as t tends to ∞. In this case one can write

$$\overline{G}(t) = e^{-\theta F(t)}, \qquad \theta > 0,\ t \geq 0, \tag{1.4}$$

where $F(t) = \Lambda(t)/\theta$ is the c.d.f. of some nonnegative random variable such that $F(0) = 0$. In what follows we will call the model given by (1.4) the bounded cumulative hazard (BCH) model.

Within the nonparametric framework it makes no difference whether representation (1.2) or (1.4) is used as a basis for the estimation of p, but the situation is not the same when $\overline{G}_0(t)$ is parametrically specified. Using definition (1.3), one can represent the survival function (1.4) in the form of formula (1.2), but both p and $\overline{G}_0$ become functions of the common parameter θ:

$$p = e^{-\theta}, \qquad \overline{G}_0(t) = \frac{e^{-\theta F(t)} - e^{-\theta}}{1 - e^{-\theta}}.$$

A similar confounding occurs when one represents (1.2) in the form of (1.4). A summary of useful formulas showing the relationship between the two models was given in Chen, Ibrahim, and Sinha (1999).

Although the models given by (1.2) and (1.4) are just two different ways of rescaling the survival function $\overline{G}(t)$, it took a long time to realize some virtues of the BCH model (which will be discussed later), which is the main subject matter of the present article. The first move in this direction was due to Haybittle (1959, 1965). The author proceeded from the observation that in some clinical data on cancer survival, an actuarial estimate of the hazard function tends to decrease exponentially with time. If the same property holds for the true hazard, expression (1.4) assumes the form:

$$\overline{G}(t) = \exp\{-\theta(1 - e^{-\zeta t})\}, \qquad \zeta > 0. \tag{1.5}$$

A comprehensive treatment of this model was given by Cantor and Shuster (1992) who used the following parameterization:

$$\overline{G}(t) = \exp\left\{\frac{\beta}{\zeta}(1 - e^{\zeta t})\right\}, \qquad \zeta \neq 0, \tag{1.6}$$

where $\beta > 0$, but ζ may take either sign. If $\zeta > 0$ the survival function (1.6) corresponds to the proper Gompertz distribution, but if $\zeta < 0$ the distribution is improper with the surviving function equaling $e^{\beta/\zeta}$. For this reason the distribution given by (1.6) can be called a generalized Gompertz distribution. In more recent work Cantor (2001) used representation (1.6) to determine the projected variance of estimated survival probabilities in clinical trials. A generalization of the model (1.6) was proposed by Cantor (1997) who suggested approximating a log hazard by a polynomial to obtain an estimate of the cure rate based on formula (1.1).

Clearly, the Gompertz-like model given by formula (1.5) is a special case of formula (1.4) with the function $F(t)$ specified by an exponential c.d.f. with parameter ζ. The representation (1.4) was first introduced in an article of Yakovlev et al. (1993) and discussed later as an alternative to the mixture model by Yakovlev (1994). Interestingly enough, Yakovlev et al. (1993) proceeded from purely biological considerations; the idea of imposing a constraint on the behavior of the hazard function was introduced later in Tsodikov, Loeffler, and Yakovlev (1998b). In fact, the authors proposed a simple mechanistically motivated model of tumor recurrence yielding an improper survival time distribution. Under this model the probability of tumor cure is defined as the probability of no clonogenic tumor cells (clonogens) surviving by the end of treatment.

The cell is called clonogenic if it is capable of producing a cell clone, that is, a group of cells that have this cell as their common parent. There is biological evidence that the majority of recurrent tumors are clonal in origin; that is, they arise from a single progenitor cell. In cancer studies the primary endpoint is conventionally the time to failure, referred to as survival time or failure time, X, with the event of failure being either tumor recurrence (disease-free survival) or death caused by the cancer under study (cancer-specific survival). According to the *clonal model* of posttreatment tumor development proposed by Yakovlev et al. (1993), a recurrent tumor arises from a single clonogenic cell. Every surviving clonogen can be characterized by a latent time (termed the *progression time*) during which it could potentially propagate into an overt tumor. Let M be the number of clonogens remaining in a treated tumor and φ_M be its probability generating function (p.g.f.). Assuming that progression times for surviving clonogens are independent and identically distributed (i.i.d.) with a common survival function $\overline{F}$, one can find easily that $\overline{G}(t) = \varphi_M(\overline{F}(t))$, $t \geq 0$. This formula suggests that knowledge of the entire distribution of the number of surviving clonogens is critical for developing biologically motivated survival models with cure. In particular, suppose that M is Poisson distributed. Then the survival function $\overline{G}(t)$ is given by the BCH model (1.4) with the parameter θ interpreted as the mean number of surviving clonogens.

A more general mechanistic model was considered by Hanin (2001). Suppose a tumor initially comprising a nonrandom number i of clonogenic cells is exposed to a fractionated radiation schedule consisting of n instantaneously delivered equal doses D separated by equal time intervals τ. It is assumed that every cell survives each exposure to the dose D with the same probability $s = s(D)$, given that it survived the previous exposures, and independently of other cells. It is also assumed that the death of irradiated tumor cells is effectively instantaneous. Assuming, in addition, that tumor growth kinetics between radiation exposures can be modeled as a homogeneous birth-and-death Markov process with birth rate $\lambda > 0$ and spontaneous death rate $\nu \geq 0$, Hanin (2001) derived an explicit formula for the distribution of the random variable (r.v.) M; the latter turned out to be a generalized negative binomial distribution. The corresponding p.g.f. is of the form (Hanin 2001):

$$\varphi_M(u) = \left(\frac{a - bu}{c - du}\right)^i, \qquad |u| \leq 1, \tag{1.7}$$

suggesting the following survival function for the time to tumor recurrence

$$\overline{G}(t) = \left(\frac{a - b\overline{F}(t)}{c - d\overline{F}(t)}\right)^i, \tag{1.8}$$

where $a = 1 - \omega - s + s\omega\mu^{n-1}$, $b = s(\omega\mu^{n-1} - 1)$, $c = 1 - \omega - s + s\mu^{n-1}$, $d = s(\mu^{n-1} - 1)$, $\alpha = e^{(\nu-\lambda)\tau}$, $\mu = s/\alpha$, $\omega = [\lambda - \nu\alpha - s(\lambda - \nu)]/[\lambda(1 - \alpha)]$, $\overline{F}(t) = 1 - F(t)$, satisfying the conditions: $\mu \neq 1$ and $\lambda \neq \nu$. The preceding formula may be taken as the starting point for the development of a new class of regression survival models; with this aim in mind, it makes sense to reduce the number of baseline parameters by enforcing the conditions: $F(0) = 0$ and $\overline{G}(0) = 1$.

Clinically detectable primary tumors are estimated to contain at least $i = 10^5$ clonogenic tumor cells ranging up to probably 10^9 cells or even more (Tucker 1999). On the other hand, the probability, s^n, for a cell to survive n fractions is expected to be very low, especially in the total dose range where cures occur frequently. This provides the rationale for exploring limiting distributions associated with model (1.8). As was shown by Hanin (2001), if n is fixed, $i \to \infty$ and $s \to 0$ in such a way that there exists a limit $A = \lim_{i\to\infty} is^n$, $0 < A < \infty$, then the distribution of the number of clonogens converges to a Poisson distribution with parameter $\theta = A/\alpha^{n-1}$. This result disproves the conjecture (Tucker, Thames, and Taylor 1990; Tucker and Taylor 1996; Tucker 1999) that, due to cell proliferation occurring between fractions of radiation, the limiting distribution of the number of surviving clonogenic cells may not be Poisson. However, the convergence to the limiting distribution is quite slow, indicating a strong point in the line of reasoning presented by these authors. The rate of convergence was evaluated numerically by Hanin, Zaider, and Yakovlev (2001) in terms of the total variation distance between the exact distribution of the number of clonogens and its Poisson limit. Another useful limiting distribution arises in the subcritical case where $\mu = s/\alpha < 1$. This condition means that the total cell loss due to both causes of cell death (radiation induced and spontaneous) prevails on average over the cell gain owing to the proliferation of tumor cells between fractions of radiation dose. An explicit expression of this distribution was derived by letting $i \to \infty$ and $n \to \infty$ so that $i\mu^n \to \gamma$, $0 < \gamma < \infty$ (Hanin et al. 2001).

The book by Yakovlev and Tsodikov (1996) presents several applications of (1.4) with a two- or three-parameter gamma distribution for the function F. The authors use the Hjort test (Hjort 1990) for goodness-of-fit testing, which is a natural choice in the parametric analysis of censored data without covariates. In a recent article, Gregori, Hanin, Luebeck, Moolgavkar, and Yakovlev (2002) adopted the Hjort goodness-of-fit test for testing several models of carcinogenesis. Methods of model diagnostics specially designed for different versions of this model have yet to be explored.

In this article, the current state of the art in methodology of statistical inference based on the BCH model is reviewed. The approach under review has emerged from cancer studies, and this fact necessarily reflects on the focus of our discussion. The history of the BCH model is relatively short, and the model has not yet enjoyed applications to other human diseases, to say nothing of nonmedical applications, such as estimation of recidivism rates. However, it should be kept in mind that other possible applications of the BCH model, including those mentioned previously, are every bit as well justified as for the binary mixture model; there is no reason for practitioners to refrain from using the BCH model just because it was originally derived in the context of cancer development.

The BCH model has the following distinct advantages:

1. It allows construction of a rich class of nonlinear transformation regression models to describe complex covariate effects. The class includes the traditional proportional hazards model as a special case; this structure is lacking in the binary mixture model. This makes the BCH model a natural tool for studying and testing departures from the proportionality of risks.

2. In some settings the BCH model provides a biologically meaningful interpretation of the results of the data analysis. This feature is especially important for combining statistical inference with other mathematical approaches to biomedical problems, for example, optimization of cancer therapy.

3. The BCH model offers certain technical advantages when developing maximum likelihood or Bayesian estimation procedures.

Before turning to the discussion of specific methods and results, one important remark is in order here. As is evident from definition (1.1), the probability of cure is essentially an asymptotic notion. However, the period of observation in actual truth is always finite, which is to say that one deals here with a problem of prediction rather than with the usual type of statistical inference. In other words, additional assumptions as to the behavior of $\overline{G}(t)$ or $\lambda(t)$ beyond the period of observation are needed. At the same time a sample of i.i.d. right-censored observations is available, and it is natural that one tries to reduce the problem of prediction to that of estimation.

2. NONPARAMETRIC METHODS FOR A HOMOGENEOUS SAMPLE

As was mentioned in Section 1, formula (1.2) is a general expression for any improper survival time distribution. Proceeding from this representation, Maller and Zhou (1996) developed a theory of nonparametric estimation of the probability p. Suppose that the survival time distribution $G(t)$ is absolutely continuous and let $t_1 < t_2 < \cdots < t_n$ be a sample (subject to right censoring) of the ordered observed failure times. Maller and Zhou (1996) suggested estimating p by

$$\hat{p}_n = \widehat{\overline{G}}_n(t_n), \qquad (2.1)$$

where $\widehat{\overline{G}}_n(t)$ is the Kaplan–Meier estimator of the underlying survival function. Then a natural estimator for the conditional survival function $\overline{G}_0(t)$ is given by

$$\widehat{\overline{G}}_{0n}(t) = \big(\widehat{\overline{G}}_n(t) - \hat{p}_n\big)/(1-\hat{p}_n), \qquad t \geq 0,\ \hat{p}_n < 1. \quad (2.2)$$

For any c.d.f. $K(t)$ define its right extreme τ_K as $\tau_K = \inf\{t \geq 0 : K(t) = 1\}$. The consistency question for the estimator $\hat{p}_n$ is settled in the following result by Maller and Zhou (1992):

Assume that censoring is independent and $0 < p \leq 1$. Let $\overline{C}(t)$ be the censoring time survival function. Let τ_H be the right extreme of $H(t) = 1 - \overline{G}(t)\overline{C}(t)$ and suppose that the c.d.f. $G(t) = 1 - \overline{G}(t)$ is continuous at τ_H in case $\tau_H < \infty$. Let $G_0(t) = 1 - \overline{G}_0(t)$ and $C(t) = 1 - \overline{C}(t)$. Then the estimator $\hat{p}_n$ is consistent if and only if

$$\tau_{G_0} \leq \tau_C. \qquad (2.3)$$

Under the same conditions the authors showed that the condition $\tau_{G_0} \leq \tau_C$ is necessary and sufficient for the convergence in probability of $\sup_{t\geq 0} |\widehat{\overline{G}}_{0n}(t) - \overline{G}_0(t)|$ to 0 as n tends to ∞. Assuming the inequality (2.3) and some additional mild conditions, Maller and Zhou also proved that, when $p < 1$, $\sqrt{n}(\hat{p}_n - p)$ is asymptotically (as $n \to \infty$) normally distributed with mean 0. Laska and Meisner (1992) showed that the estimator $\widehat{\overline{G}}_n(t_n)$ is, in fact, the generalized nonparametric maximum likelihood estimator of the proportion p in the mixture model (1.2).

As discussed previously, the inequality $\tau_{G_0} \leq \tau_C$ is both necessary and sufficient for consistency of the Kaplan–Meier estimator of the cure probability p. This inequality may be thought of as a quantification of "sufficient follow-up" (Maller and Zhou 1992). If the condition is not true, then failures may occur after the maximum follow-up period and it is not possible to determine which proportion of the late censored data has actually been cured. Maller and Zhou (1994) suggested a statistical test based on the length $t_n - t_n^*$ of the interval between the largest uncensored failure time t_n^* and the largest overall failure time t_n. Intuitively, if this interval is large, then the last failure has occurred well before the last censoring event so there has been sufficient follow-up. Maller and Zhou showed that, under appropriate regularity conditions, the estimator $\alpha_n = (1 - N_n/n)^n$ is an approximate p value for a test of $\tau_{G_0} \leq \tau_C$, and α_n converges to 0 in probability if and only if $\tau_{G_0} \leq \tau_C$. Here N_n is the number of uncensored failure times in the interval $(2t_n^* - t_n, t_n^*]$. This test estimates a significance level rather than controls for it, so that the original version of the test by Maller and Zhou suggests the hypothesis of sufficient follow-up to be rejected whenever the estimated p value exceeds a prespecified critical value, say .05. Later, after conducting computer simulations, Maller and Zhou (1996) came to the conclusion that this test was far too conservative. It is clear that a pertinent test should be based on percentiles of the sample distribution of α_n (or the closely related statistic $q_n = N_n/n$) rather than on a fixed p value. Unfortunately, the relevant sample distributions are not known even in large samples. The idea of sufficient follow-up is intuitively compelling and a search for a more general formal definition of this notion (which is not necessarily related to consistency of the corresponding nonparametric estimator) should be continued.

When proceeding from the BCH model, formula (1.1) suggests the following nonparametric estimator of the cure probability: $\tilde{p}_n = \exp\{-\tilde{\Lambda}_n(t_n)\}$, where $\tilde{\Lambda}_n(t_n)$ is the Nelson–Aalen estimator of the cumulative hazard at the point of last observation. A pertinent nonparametric estimator for $F(t)$ can be proposed in the form:

$$\widehat{F}_n(t) = \log \widehat{\overline{G}}_n(t) / \log \widehat{\overline{G}}_n(t_n), \qquad (2.4)$$

where $\widehat{\overline{G}}_n(t)$ is the Kaplan–Meier estimator for the survival function $\overline{G}(t)$.

It is important to note that the preceding nonparametric estimators yield an estimate of survival probability at the right end T of the observation period. This estimate is all we have to predict the behavior of the survival function beyond the period of observation, and all such predictions are final. If we strictly follow the formal definition of cure rate, we have to recognize that the nonparametric approach implies a straight-line extrapolation of the estimated survival function beyond the period of observation T: $\overline{G}(t)$ is set to be equal to a constant value of $\widehat{\overline{G}}_n(t_n)$ for all $t \geq T$. The same holds true for the semiparametric regression models discussed in Section 3. It should also be noted that the length T of the period of observation is implicitly involved in (2.3). Let us decompose the censoring time survival function as follows: $\overline{C}(t) = \overline{C}^*(t)[1 - I(t - T)]$, where $I(x) = 0$ for $x < 0$ and $I(x) = 1$ for $x \geq 0$. Then inequality (2.3) is replaced by $\tau_{G_0} \leq \min\{\tau_{C^*}, T\}$. The finiteness of the follow-up period is relevant to large-sample studies, where the value of T should be kept fixed when increasing the sample size; otherwise, ensuring asymptotic properties would imply infinite observation time.

Another feature inherent in the nonparametric approach has to do with the instability (high variance) of the nonparametric estimators. It is a well-known fact that the Kaplan–Meier estimator becomes highly unstable for t close to the end of an observation period in the presence of heavy censoring (Pepe and Fleming 1989; Cantor and Shuster 1992; Tsodikov 2001), which may have a very significant effect on the accuracy of cure rate estimation. The method for testing for sufficient follow-up, proposed by Maller and Zhou (1994, 1996), suffers from the same kind of instability as well. The nonparametric estimator $\widehat{F}_n(t) = \log \widehat{\overline{G}}_n(t)/\log \widehat{\overline{G}}_n(t_n)$ is particularly sensitive to variations of $\widehat{\overline{G}}_n(t_n)$ in the denominator. Therefore, it is not recommended to use it in its original form. An improved estimator was proposed in Tsodikov (2001). This is a two-stage estimator. First, $\overline{G}_n(t_n)$ is estimated parametrically. Then $\overline{G}(t)$ is estimated nonparametrically with $\overline{G}_n(t_n)$ constrained to be equal to the corresponding value of the parametric estimate.

3. SEMIPARAMETRIC REGRESSION SURVIVAL MODELS WITH CURE

3.1 Mixture Models and Generalizations

So far we have followed two distinct lines of reasoning in the discussion of cure models: statistical and mechanistic. We now consider more formal relationships between the two aspects of the problem.

A cure model can be formulated by making assumptions about either the hazard or the survival function. For example, making the assumption on the bounded cumulative hazard in a proportional hazards (PH) model, we obtain the so-called improper PH model

$$\overline{G}(t|\boldsymbol{\beta},\mathbf{z}) = \exp\{-\theta(\boldsymbol{\beta},\mathbf{z})F(t)\}, \qquad (3.1)$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients, $\mathbf{z}$ is a vector of covariates, and $\theta(\boldsymbol{\beta},\mathbf{z})$ is a known function relating $\boldsymbol{\beta}$ to $\mathbf{z}$. In what follows we allow for more than one predictor, and for an arbitrary parameterization of regression predictors. However, in the examples the most common exponential parameterization is used, in which $\theta(\boldsymbol{\beta},\mathbf{z}) = \exp\{\boldsymbol{\beta}'z\}$, where $\boldsymbol{\beta}'$ is the transposed $\boldsymbol{\beta}$ vector. The component β_0 of the vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots)'$ is the intercept term; therefore, $z_0 = 1$ in the vector of covariates $\mathbf{z} = (z_0, z_1, \dots)$.

A general class of semiparametric regression models, nonlinear transformation models (NTM), was proposed in Tsodikov (2002, 2003):

$$\overline{G}(t|z) = \gamma\big(\overline{F}(t)|\boldsymbol{\beta},\mathbf{z}\big), \qquad (3.2)$$

where $\gamma(x|\boldsymbol{\beta},\mathbf{z})$ is some parametrically specified cumulative distribution function in x with support on [0, 1]. Although our discussion allows for any parameterization of γ in terms of $\boldsymbol{\beta}$ and $\mathbf{z}$, in the examples we assume that γ is parameterized through a set of parameters/predictors $\theta, \eta, \dots$, where each predictor is further parameterized using generally different sets of regression coefficients $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots$, so that $\theta = \exp\{\boldsymbol{\beta}_1'\mathbf{z}\}$, $\eta = \exp\{\boldsymbol{\beta}_2'\mathbf{z}\}$. The linear transformation models considered by Cheng, Wei and Ying (1995) represent a subclass of NTM with

$$\gamma(x|\boldsymbol{\beta},\mathbf{z}) = p[\log\theta(\boldsymbol{\beta},\mathbf{z}) + q(x)], \qquad (3.3)$$

where p is a tail function (=1− c.d.f.) and q is an inverse of a tail function (not necessarily that of p). Cure models represented by the NTM class were introduced in Tsodikov (2002). To make (3.2) a cure model, the following assumptions are made to enforce the limit $\overline{G}(t|\boldsymbol{\beta},\mathbf{z}) \to p > 0$:

$$\gamma(0|\boldsymbol{\beta},\mathbf{z}) \equiv h(\boldsymbol{\beta},\mathbf{z}) > 0, \qquad \lim_{t\to\infty} \overline{F}(t) = 0. \qquad (3.4)$$

The restriction $\lim_{t\to\infty} \overline{F}(t) = 0$ proposed by Taylor (1995) in the context of the two-component mixture model removes the overparameterization of the description of the baseline cure rate through F and h and the associated estimation and convergence problems for the fitting algorithms. Following the restriction, the estimate $\widehat{\overline{F}}$ is assumed to satisfy $\widehat{\overline{F}}$(last failure) = 0. This restriction is also necessary for separation of the long- and short-term covariate effects on survival. The preceding restriction should be observed when parameterizing the model, as discussed in the following examples.

Alternatively, a cure model can be formulated as a two-stage model. First, an unobservable random variable M is postulated with distribution $p(m|\boldsymbol{\beta},\mathbf{z})$ that depends on covariates. Second, the observed survival function is formed as

$$\overline{G}(t|\boldsymbol{\beta},\mathbf{z}) = \mathrm{E}\big[\overline{F}(t)^M|\mathbf{z}\big], \qquad (3.5)$$

where the expectation is taken with respect to M, given $\mathbf{z}$, and the distribution of M depends on $\boldsymbol{\beta}$ and $\mathbf{z}$. The model (3.5) represents a generalized PH frailty model, which we simply call the PH mixture model, introduced in Tsodikov (2002, 2003), where M is assumed to be an arbitrary nonnegative random variable. In this context M can be interpreted as a missing covariate in a PH model. Obviously, the usual PH frailty model

$$\overline{G}(t|\boldsymbol{\beta},\mathbf{z}) = \mathrm{E}\big[\overline{F}(t)^{U\theta(\boldsymbol{\beta},\mathbf{z})}|\mathbf{z}\big], \qquad (3.6)$$

where the distribution of U is independent of covariates, is a particular case of (3.5) with $M = U\theta(\boldsymbol{\beta},\mathbf{z})$. Missing variables U dependent on covariates have been considered by several authors (e.g., Wassel and Moeschberger 1993) within the shared PH frailty framework. It should be noted that the PH mixture model (3.5), although at least as general as (3.6) with $U = U(\boldsymbol{\beta},\mathbf{z})$, leads to a computationally efficient estimation procedure that we consider in the next section. Because all models considered previously in this article are particular cases of (3.5), an estimation procedure designed for the PH mixture model, or for the NTM subclass (3.2) restricted to (3.4), represents a universal tool for statistical inference with such models.

The discrete mixture model [i.e., (3.5) with a discrete random variable M], which was used in particular forms as a mechanistic model in Section 1, can be linked to the NTM class. Observe that the p.g.f. of a discrete random variable M

$$\varphi_M(x) = \sum_{m=0}^{\infty} p_m x^m$$

is a distribution function in x with the support [0, 1]. Indeed, because the p_m are nonnegative, $\varphi_M(x)$ is increasing in x. Also,

$\varphi_M(1) = \sum_{m=0}^{\infty} p_m = 1$. If M is discrete and depends on covariates, (3.5) can be rewritten as an NTM with

$$\gamma(x|z) = \varphi_M(x|z).$$

With the discrete mixture model, the probability of cure is given by $\gamma(0|z) = \varphi_M(0|\boldsymbol{\beta}, \mathbf{z}) = p_0(\boldsymbol{\beta}, \mathbf{z})$.

Alternatively, we may want to write an NTM model as a discrete mixture model. It should be noted that NTM is a wider class than that of mixture models. For an NTM to be a discrete mixture, we may require that $\gamma(x|\boldsymbol{\beta}, \mathbf{z})$ be an analytic function of x on [0, 1]. As such, it can be expanded in a power series about $x = 0$. Additionally, the coefficients of the corresponding power series must be positive. Should this be the case, they can be interpreted as probabilities, and γ as a generating function of some discrete random variable. If γ is a mixture model (discrete or continuous), then $\psi(\lambda) = \gamma(e^{-\lambda})$ is the Laplace transform of the distribution of the mixing variable M. Generally, a necessary and sufficient condition for $\gamma(x|\cdot)$ to be a mixture model is that ψ be a completely monotonic function as given by the Bernstein theorem (see Feller 1971). A function $\psi(\lambda)$ is called completely monotonic if all of the derivatives $\psi^{(n)}$ exist and $(-1)^n \psi^{(n)}(\lambda) \geq 0$, $\lambda > 0$.

We can now represent the cure models discussed earlier in this article as members of the mixture–NTM model family. It should be stressed that such a representation is not unique.

The improper PH model (3.1) corresponds to

$$\gamma(x|\boldsymbol{\beta}, \mathbf{z}) = \exp\{-\theta(\boldsymbol{\beta}, \mathbf{z})(1 - x)\}. \qquad (3.7)$$

This is the generating function of the Poisson distribution, and by expanding γ about $x = 0$, we get a power series with Poisson probabilities as coefficients. This gives us the mechanistic interpretation of the improper PH model discussed in Section 1.

Consider an extension of the improper PH model allowing for dissimilar covariate effects on long- and short-term survival. To construct an extended hazard model (the term was introduced by Etezadi-Amoli and Ciampi 1987), we employ the fact that $\overline{F}$ is a survival function. Incorporating covariates into $\overline{F}$, we can add a short-term effect to the improper PH model. The class of extended hazard models

$$\overline{G}(t|\boldsymbol{\beta}, \mathbf{z}) = \exp\{-\theta(\boldsymbol{\beta}_1, \mathbf{z})[1 - \tilde{\gamma}(\overline{F}(t)|\boldsymbol{\beta}_2, z)]\}, \qquad (3.8)$$

where $\tilde{\gamma}$ is an NTM and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, $\boldsymbol{\beta}_i = (\beta_{i0}, \beta_{i1}, \ldots)$, $i = 1, 2$, was introduced in Tsodikov (2002). If $\tilde{\gamma}$ is a mixture model itself, then (3.8) is also a mixture model with $\gamma = \exp\{-\theta[1 - \tilde{\gamma}]\}$. The mixing variable M that generates the family (3.8) has a well-defined structure of a compound variable

$$M = \sum_{k=1}^{\nu} \delta_k, \qquad (3.9)$$

where ν is a Poisson random variable, $\sum_1^0 = 0$, and the δ_k are i.i.d. copies of a random variable δ with Laplace transform $\tilde{\gamma}(e^{-\lambda})$. The binary distribution for ν gives rise to the two-component mixture class of models

$$\overline{G}(t|\boldsymbol{\beta}, \mathbf{z}) = p(\boldsymbol{\beta}_1, \mathbf{z}) + \bar{p}(\boldsymbol{\beta}_1, \mathbf{z})\tilde{\gamma}[\overline{F}(t)|\boldsymbol{\beta}_2, \mathbf{z}], \qquad \bar{p} = 1 - p. \qquad (3.10)$$

Kuk and Chen (1992) proposed a regression model in the form of (3.10) with p being a logistic regression and $\tilde{\gamma}$ a proportional hazards regression. This model was further studied by Sy and Taylor (2000) and Peng and Dear (2000). The Poisson distribution for ν leads to the bounded hazard family of mixture models. For example, let δ be degenerate (nonrandom): $\Pr(\delta = \eta(\boldsymbol{\beta}_2, \mathbf{z})) = 1$. Then $M = \eta(\boldsymbol{\beta}_2, \mathbf{z})\nu$, where ν is Poisson with expectation $\theta(\boldsymbol{\beta}_1, \mathbf{z})$. Then we have the PHPH model

$$\overline{G}(t|\boldsymbol{\beta}, \mathbf{z}) = \exp[-\theta(\boldsymbol{\beta}_1, \mathbf{z})\{1 - \overline{F}(t)^{\eta(\boldsymbol{\beta}_2, \mathbf{z})}\}]. \qquad (3.11)$$

The model (3.11) was proposed by Broët et al. (2001) in the context of two sample score tests for long- and short-term covariate effects.

In the preceding cure models, restriction (3.4) needs to be observed. In models (3.8) considered so far, one predictor (η) was used to describe the short-term effect. To avoid overparameterization of the model, the intercept component of $\boldsymbol{\beta}_2$ has to be fixed or removed, for example, by setting $\beta_{20} = 0$. With this coding of the model, the intercept term in the first predictor (θ), β_{10}, corresponds to the baseline log cure rate. With an exponential parameterization of the predictors, the baseline survival function takes the form $\overline{G}_b(t|\boldsymbol{\beta}_0, \mathbf{z}) = \exp\{-\exp(\beta_{10})[1 - \tilde{\gamma}(\overline{F}|\mathbf{0})]\}$, where $\mathbf{0} = (0, 0, \ldots)'$ and $\boldsymbol{\beta}_0 = ((\beta_{10}, 0, 0, \ldots), \mathbf{0})'$. As in the Cox model (Cox 1972), the regression coefficients β_{ij}, $i = 1, 2$, $j = 1, 2, \ldots$, correspond to the relative effects of the covariates z_{ij} on the long- and short-term predictor, $i = 1, 2$, respectively. For example, with the PHPH model, β_{ij}, $i = 1, 2$, $j = 1, 2, \ldots$, is the log of the hazard ratio in the two PH models of which the PHPH model is composed, corresponding to the long- and short-term effect, $i = 1, 2$, respectively.

3.2 Estimation Procedures

The issue of the potentially infinite dimension has been the most critical deterrent to the use of maximum likelihood estimation (MLE) in semiparametric regression models. Methods based on the partial likelihood are specific to the proportional hazards model and do not extend to other models. The Newton–Raphson procedure requires taking the inverse of the information matrix, which gets computationally prohibitive and unstable with increasing dimension. Characterizing the future directions in survival analysis in their editorial in *Biometrics*, Fleming and Lin (2000) pointed out that "it would be highly useful to develop efficient and reliable numerical algorithms for the semiparametric estimation. . . ." In the following sections we review some recent approaches to the problem.

Introduce a set of times t_i, $i = 1, \ldots, n$, arranged in increasing order, with $t_{n+1} = \infty$. Associated with each t_i is a set of individuals $\mathcal{D}_i$ with covariates $\mathbf{z}_{ij}$, $j \in \mathcal{D}_i$, who fail at t_i, and a similar set of individuals $\mathcal{C}_i$ with covariates $\mathbf{z}_{ij}$, $j \in \mathcal{C}_i$, who are censored at t_i. For any function $A(t)$ let $A_i = A(t_i)$ and $\Delta A_i = |A_i - A_{i-1}|$. The generalized log-likelihood for an NT model (3.2) can be written as

$$\ell = \sum_{i=1}^{n} \left\{ \sum_{j \in \mathcal{D}_i} \log[\gamma(\overline{F}_{i-1}|\mathbf{z}_{ij}) - \gamma(\overline{F}_i|\mathbf{z}_{ij})] + \sum_{j \in \mathcal{C}_i} \log \gamma(\overline{F}_i|\mathbf{z}_{ij}) \right\}. \qquad (3.12)$$

3.2.1 ConventionalMaximizationMethods. Considermaximization of the log-likelihood function $\ell(\mathbf{x})$ with respect to the joint vector of parameters $\mathbf{x}$, representing regression coefficients $\boldsymbol{\beta}$ and the nonparametrically specified function F. The vector $\Delta\mathbf{H}$ represents a set of jumps ΔH_i of the cumulative hazard function H, which can be used to parameterize F so that $\mathbf{x} = (\boldsymbol{\beta}, \Delta\mathbf{H})$. Consider a so-called direction sets method (Press, Flannery, Teukolsky, and Vetterling 1994) constructed as follows:

1. Given the current iteration vector $\mathbf{x}^{(k)}$, find the search direction $s^{(k)}$.
2. Maximize $f(\mathbf{x}^{(k)} + ys^{(k)})$ with respect to the scalar y.
3. Set $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + y^*s^{(k)}$, where y^* is the solution at the previous step.
4. Set $k = k + 1$ and continue to step 1.

The Newton–Raphson method is one example of the direction sets methods. To avoid taking the inverse of the full-model information matrix, the search direction s can be specified according to the Powell method, which uses multiple line maximizations in one dimension to construct a set of conjugate directions (Press et al.,1994).

3.2.2 Profile LikelihoodApproach. Assume that we have a method to obtain the global maximum of ℓ with respect to F, given $\boldsymbol{\beta}$. We discuss two such methods in the next two sections. We may write the profile log-likelihood as an implicit function of $\boldsymbol{\beta}$:

$$\ell_{\text{pr}}(\boldsymbol{\beta}) = \ell\big(\boldsymbol{\beta}, F^*(\boldsymbol{\beta})\big), \tag{3.13}$$

where F^* is the solution of the problem

$$\ell_{\text{pr}}(\boldsymbol{\beta}) = \max_{F \in \mathcal{F}'} \ell, \tag{3.14}$$

where $\mathcal{F}'$ is the class of proper discrete survival functions. A profile algorithm is a straightforward nested procedure:

- Maximize $\ell_{\text{pr}}(\boldsymbol{\beta})$ by a conventional nonlinear programming method (e.g., a direction sets method).
- For any $\boldsymbol{\beta}$ as demanded in the preceding maximization procedure, solve the problem (3.14) with specified tolerance.

Inference based on the profile likelihood is similar to that based on the partial likelihood for the PH model. However, the classical theory of MLE does not apply to infinite dimensions. Important results have been obtained regarding theoretical justifications for the nonparametric maximum likelihood estimation (NPMLE) method and the profile likelihood for semiparametric models (Murphy 2000). It was shown that profile likelihoods with nuisance parameters estimated out behave like ordinary likelihoods under some conditions. In particular, these results apply to the PH model, the proportional odds (PO) model, the PH frailty model, and presumably to many other models.

3.2.3 Restricted Nonparametric Maximum Likelihood Estimation. Tsodikov (2002) developed the so-called restricted nonparametric maximum likelihood estimation (RNPMLE) algorithm. The RNPMLE method is based on a recurrent structure of the score equations with respect to the nonparametric part of the model ($\overline{F}$). Earlier variations of the procedure were developed for the PH model (Tsodikov 1998a) and for the proportional odds (PO) model (Tsodikov, Hasenclever, and Loeffler 1998a) and were subsequently generalized for NT cure models (Tsodikov 2002). Although quite computationally intensive, the RNPMLE method is numerically stable and it allows us to avoid taking inverses or approximating the information matrix of the full model.

According to the nonparametric maximum likelihood method, the model can be fitted by maximizing the generalized log-likelihood (3.12) with respect to the regression parameters $\boldsymbol{\beta}$ and the unspecified survival function $\overline{F}$ (or the corresponding distribution function F). In doing so, the log-likelihood ℓ is maximized for a step function F with steps at the times of failures.

As discussed earlier we impose the restriction $F(t_n) = 1$. To maximize (3.12) under the restriction $F(t_n) = 1$, we use the method of Lagrange multipliers and obtain the system of score equations

$$\frac{\partial \ell(\boldsymbol{\beta}, F)}{\partial \boldsymbol{\beta}} = 0, \tag{3.15}$$

$$\frac{\partial \ell(\boldsymbol{\beta}, F)}{\partial F_i} = 0, \qquad i = 1, \dots, n, \tag{3.16}$$

$$F_n = 1. \tag{3.17}$$

Inference procedures for the regression coefficients $\boldsymbol{\beta}$ can be obtained from the profile log-likelihood (3.14). The profile log-likelihoodis obtained by solving the score equations (3.16) and (3.17) for F simultaneously, given $\boldsymbol{\beta}$.

In the RNPMLE method the system of score equations (3.16) is partitioned into a subsystem of the form

$$\frac{\partial \ell}{\partial F_i} = \psi_i(F_{i-1}, F_i, F_{i+1}) = 0 \tag{3.18}$$

that can be solved recurrently for $\overline{F}_2, \dots, \overline{F}_n$, given $F_0 = 0$ and F_1. Finally, the equation $\varphi(\overline{F}_1) = 1$ is solved, where φ is defined as F_n generated recurrently using the subsystem (3.18). The RNPMLE method yields point estimates and confidence intervals for the regression coefficients and the function F (Tsodikov 2002).

3.2.4 EM-Based Methods. The EM algorithm used to fit shared frailty models in survival analysis (Nielsen, Gill, Andersen, and Sørensen 1992) handles H in a numerically efficient way. This is made possible as the M step reduces to the PH model. Estimates are obtained iteratively by maximizing the partial likelihood and computing the Nelson–Aalen–Breslow estimator (Andersen, Borgan, Gill, and Keiding 1993) for the cumulative hazard H. Similar algorithms have been used to fit the two-component mixture model, given by Taylor (1995), Sy and Taylor (2000, 2001), and Peng and Dear (2000).

Recently, Tsodikov (2003) generalized the EM algorithm for the frailty model into a universal "distribution-free" procedure applicable to the general PH mixture model (3.5) and NTM class (3.2). This family of algorithms (quasi-EM, QEM) is a subclass of the so-called MM algorithms based on surrogate objective functions (Lange, Hunter, and Yang 2000). Broadly defined, an MM algorithm substitutes a computationally simpler surrogate objective function for the target function on each step of the procedure (similar to the E step of the EM). Maximizing the surrogate objective function drives the target function in the desired direction. Thus, a difficult maximization problem is replaced by a series of simpler ones. Unfortunately, there is no universal recipe on how to find an appropriate surrogate objective function. The idea of the QEM is to obtain a surrogate objective function by generalizing the one inherent in the EM procedure so that it does not depend on the missing data formulation of the model. Practically, the QEM approach for semiparametric models is based on a derivation of the "distribution-free" E step for the EM algorithm, constructed for the PH mixture model (3.5). More formally, we consider the QEM procedure for cure models. The following statement is the key result underlying the QEM construction.

Proposition 3.1 (Tsodikov 2003). Let τ be the observed event time under independent censoring and let c be the observed censoring indicator ($c = 1$, if a failure, and $c = 0$, otherwise). Under the PH mixture model,

$$G(t|\cdot) = \gamma\big[\overline{F}(t)|\cdot\big] = \mathrm{E}\big\{\big[\overline{F}(t)\big]^{M(\cdot)}\big\},$$

and if $F'(\tau) > 0$, the conditional expectation of M, given the observed event, is given by

$$\mathrm{E}\big\{M(\cdot)|\cdot,\tau,c\big\} = \Theta\big[\overline{F}(\tau)|\cdot,c\big],$$

where

$$\Theta\,[x|\cdot,c] = c + x\frac{\gamma^{(c+1)}(x|\cdot)}{\gamma^{(c)}(x|\cdot)}, \tag{3.19}$$

where $\gamma^{(c)}(x|\cdot) = \partial^c\gamma(x|\cdot)/\partial x^c$, $c = 0, 1, \ldots$, $\gamma^{(0)}(x|\cdot) = \gamma(x|\cdot)$.

The preceding result indicates that the E step can be constructed using the first two derivatives of the NTM-generating function γ without any knowledge or even existence of the mixing random variable M. Specification of an algorithm for a particular model requires evaluation of Θ for the model. For example, for the models discussed in the previous section, we have

- Improper PH model: From (3.1) or (3.7), $\Theta(x|\cdot,c) = c + \theta(\cdot)x$.
- PHPH model: From (3.11), $\Theta(x|\cdot,c) = \theta(\cdot)\eta(\cdot)x^{\eta(\cdot)} + c\eta(\cdot)$.

Cure models require a correction of Θ at the last observation [see (3.21)]. Let $c_{ij} = 1$, $j \in \mathcal{D}_i$, and $c_{ij} = 0$, $j \in \mathcal{C}_i$. Let τ_k, $k = 1, \ldots, K$, be the time points in ascending order where failures occur ($\mathcal{D}$ is not empty). Denote by r_k the rank of τ_k in the set $\{t_i\}$. The ranks r_k locate the points τ_k on the set $\{t_i : \tau_k = t_{r_k}\}$. By definition, we set $\tau_{K+1} = n + 1$ and $\tau_0 = 0$. Imputation of the mixing variable M using (3.19) results in the score equations for the cumulative hazard H corresponding to F,

$$\Delta H_k = \frac{D_k}{\sum_{ij\in\mathcal{R}_k}\Theta(\overline{F}_i|\boldsymbol{\beta},\mathbf{z}_{ij})}, \tag{3.20}$$

where $\mathcal{R}_k = \bigcup_{i=k}^{n}\{\mathcal{C}_i \cup \mathcal{D}_i\}$ is the risk set at t_k, D_k is the number of failures at t_k, and

$$\Theta(\overline{F}_i|\cdot_{ij}) = \begin{cases} \dfrac{\gamma'(\overline{F}_i|\cdot_{ij})}{\gamma(\overline{F}_i|\cdot_{ij})}\overline{F}_i, & j \in \mathcal{C}_i,\ i \le r_K - 1, \\ 1 + \dfrac{\gamma''(\overline{F}_i|\cdot_{ij})}{\gamma'(\overline{F}_i|\cdot_{ij})}\overline{F}_i, & j \in \mathcal{D}_i,\ i \le r_{K-1}, \\ \dfrac{\gamma'(\overline{F}_{r_{K-1}}|\cdot_{ij})}{\gamma(\overline{F}_{r_{K-1}}|\cdot_{ij}) - \gamma(0|\cdot_{ij})}\overline{F}_{r_{K-1}}, & j \in \mathcal{D}_{r_K},\ i = r_K, \\ 0, & j \in \mathcal{C}_i,\ i \ge r_K, \end{cases} \tag{3.21}$$

where $(\cdot)_{ij}$ stands for $(\boldsymbol{\beta}, \mathbf{z}_{ij})$. The fact that the contributions of the observations ij to the likelihood as well as Θ at, or after the last failure ($i \ge r_K$), are different from their counterparts before the last failure ($i \le r_K - 1$) is due to the restriction $\overline{F}_i = 0$, $i = r_K, \ldots, n$, representing the fact that $\overline{F}$ is a proper survival function. This distinction is not made if the model is a general NTM with an unrestricted F.

It is interesting to note that the score equations have the form of the Nelson–Aalen–Breslow estimator for the PH model with the usual predictor replaced by Θ. Because Θ depends on F, an iteration procedure is needed to satisfy (3.20). Given $\boldsymbol{\beta}$, iterations with respect to F can be carried out as follows:

- At the kth iteration $\overline{F}^{(k)}$, compute $\Theta^{(k)}$ for each subject.
- Update $F^{(k)}$ by $F^{(k+1)}$ using the Nelson–Aalen–Breslow estimator (3.20).

It can be shown (Tsodikov 2003) that if $\Theta(x|\cdot)$ is a nondecreasing function of x, or if γ is a PH mixture model (a stronger assumption), then each iteration described previously improves the likelihood. Also, this assumption makes the preceding procedure a member of the MM family (Lange et al. 2000), and the convergence properties of the algorithm follow from the general MM theory.

There are many ways to build a particular model fitting algorithm based on the principles described previously, and this depends on how the maximization with respect to $\boldsymbol{\beta}$ is incorporated into the procedure. One way would be to use partial likelihood and a setup similar to the EM algorithm for frailty models (Nielsen et al. 1992). Such an EM algorithm was used in Sy and Taylor (2000) and Peng and Dear (2000) to fit the two-component mixture model. As simple as the setup of the procedure with a binary distribution for M might be, its generalization for a wider family of PH mixture models or NTM's is problematic. The key to a straightforward setup of the procedure is not to use the EM principle on $\boldsymbol{\beta}$, as it would require working with the distribution of M.

We find the profile version of the QEM algorithm particularly easy and straightforward to work with. The profile QEM algorithm outperformed conventional maximization procedures (Sec. 3.2.1) and the frailty EM algorithm based on partial likelihood when used to fit a semiparametric proportional odds (PO) model. Also, it was shown to be faster than a directions set method used to fit a parametric PO model.

3.3 Example: Prostate Cancer Data

The study was carried out using data on 1,100 patients with clinically localized prostate cancer who were treated with three-dimensional conformal radiation therapy at the Memorial Sloan-Kettering Cancer Center (Zaider et al. 2001). The patients were stratified by radiation dose (group 1, <67.5 Gy; group 2, 67.5–72.5 Gy; group 3, 72.5–77.5 Gy; group 4, 77.5–87.5 Gy) and prognosis category [favorable, intermediate, and unfavorable as defined by pretreatment prostate specific antigen (PSA) measurement and Gleason prognostic score]. A relapse was recorded when tumor recurrence was diagnosed or when three successive PSA elevations were observed from a posttreatment nadir PSA level. PSA relapse-free survival was used as the primary endpoint. There were no failures observed in patients with favorable prognosis in dose group 4. For that reason this group of patients is not included in the data analysis.

We applied the PHPH extended hazard model (3.11) with $\mathbf{z}$ specified by indicator dummy variables that code the four dose groups and the three prognostic categories. In the unrestricted model dose and summary prognosis may each have an effect on long-term survival through $\theta(z)$ and on short-term survival through $\eta(z)$. The hypothesis of no short-term effect of prognostic category (proportional hazards with respect to prognosis) is not rejected ($p = .49$) by the likelihood ratio test. This observation justifies the use of the PH model for the effect of the prognostic category. However, the total radiation dose appears to have a significant ($p < .0001$) short-term effect on survival of patients with prostate cancer, indicating nonproportionality of the effect of dose. Resorting to our mechanistic interpretation of (3.11), we may speculate that the progression time distribution varies with radiation dose while being essentially the same for all prognostic categories. Semiparametric estimates of the mean values for the distributions $1 - \gamma(\overline{F}|z)$ are equal to 1,279, 867, 803, and 498 days for dose groups 1, 2, 3, and 4, respectively. Estimates of the hazard ratio for the short-term effect of each dose group,

$$\frac{\log \gamma(\overline{F}|\text{dose group } i)}{\log \gamma(\overline{F}|\text{dose group } 4)} = B(\text{dose group } i),$$

show a monotonically increasing (with dose) short-term risk: .188 with confidence limits (.038, 1.538) for dose group 1, .415 (.065, .765) for dose group 2, .474 (.024, 1.224) for dose group 3, and 1.000 for dose group 4.

Semiparametric estimates of the probability of cure are given in Table 1. For comparison, we also present parametric estimates obtained with F specified by a two-parameter gamma distribution. It is shown in Table 1 that the two estimates appear to be quite close to each other. This analysis indicates that, in terms of cure rate, dose escalation has a significant positive effect only in the intermediate and unfavorable groups. It is also found that progression time is inversely proportional to dose, which means that patients recurring in higher dose groups have shorter recurrence times, yet these groups have better long-term survival. One possible explanation for this seemingly illogical observation lies in the fact that less aggressive tumors potentially recurring after a long period of time are cured by higher doses and do not contribute to the observed time pattern of tumor relapse. As a result tumors in higher dose groups are less likely to recur; however, if they do, they tend to recur earlier.

Table 1. Estimates of the Probability of Cure (semiparametric/parametric) and 95% Semiparametric Likelihood Ratio Confidence Intervals (in parentheses) as Estimated Using Multivariate Semiparametric Regression Analysis Based on (3.11).

Prognostic	*Dose group*			
category	*1*	*2*	*3*	*4*
Favorable	.78/.80	.79/.74	.88/.87	1.00*
	(.55,.95)	(.68,.92)	(.80,.97)	–
Intermediate	.37/.25	.53/.51	.67/.58	.79/.74
	(.21,.55)	(.41,.64)	(.58,.78)	(.68,.87)
Unfavorable	.02/.00	.35/.27	.46/.33	.60/.64
	(.00,.11)	(.25,.45)	(.38,.56)	(.45,.74)

*The estimate of the probability of cure is set equal to 1 because no failures are observed in this group of patients.

4. BAYESIAN INFERENCE

4.1 Parametric Cure Rate Model

Ibrahim, Chen, and Sinha (2001a) presented a comprehensive treatment of Bayesian approaches for the cure rate model

$$\overline{G}(t) = \exp\{-\theta F(t)\}. \tag{4.1}$$

We review some of their work here. We let the covariates depend on θ through the relationship $\theta = \exp(\mathbf{x}'\boldsymbol{\beta})$, where $\mathbf{x}$ is a $p \times 1$ vector of covariates and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is a $p \times 1$ vector of regression coefficients. Proceeding from the biological interpretation discussed in Section 1, we can now construct the likelihood function (see Chen et al. 1999; Ibrahim et al. 2001a) in a typical setting. Suppose we have n subjects and let N_i denote the number of surviving clonogenic tumor cells for the ith subject. Further, we assume that the N_i's are i.i.d. Poisson random variables with mean θ, $i = 1, 2, \ldots, n$. We emphasize here that the N_i's are not observed and can be viewed as latent variables in the model formulation. Further, suppose $Z_{i1}, Z_{i2}, \ldots, Z_{i,N_i}$ are the i.i.d. progression times for the N_i cells for the ith subject, which are unobserved, and all have proper cumulative distribution function $F(\cdot)$, $i = 1, 2, \ldots, n$. We denote the indexing parameter (possibly vector valued) by $\boldsymbol{\psi}$, and thus write $F(\cdot|\boldsymbol{\psi})$ and $\overline{F}(\cdot|\boldsymbol{\psi}) = 1 - F(\cdot|\boldsymbol{\psi})$. For example, if $F(\cdot|\boldsymbol{\psi})$ corresponds to a Weibull distribution, then $\boldsymbol{\psi} = (\alpha, \lambda)'$, where α is the shape parameter and λ is the scale parameter. Let y_i denote the survival time for subject i, which may be right censored, and let ν_i denote the censoring indicator, which equals 1 if y_i is a failure time and 0 if it is right censored. The observed data are $D_{\text{obs}} = (n, \mathbf{y}, \boldsymbol{\nu})$, where $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ and $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_n)'$. Also, let $\mathbf{N} = (N_1, N_2, \ldots, N_n)'$. The complete data are given by $D = (n, \mathbf{y}, \boldsymbol{\nu}, \mathbf{N})$, where $\mathbf{N}$ is an unobserved vector of latent variables. Throughout the remainder of this section, we will assume a Weibull density for $f(y_i|\boldsymbol{\psi})$, so that $f(y|\boldsymbol{\psi}) = \alpha y^{\alpha-1}\exp\{\lambda - y^{\alpha}\exp(\lambda)\}$. When covariates are included we have a different cure rate parameter, θ_i, for each subject, $i = 1, 2, \ldots, n$. Let $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip})$ denote the $p \times 1$ vector of covariates for the ith subject. We relate θ to the covariates by $\theta_i \equiv \theta(\mathbf{x}_i'\boldsymbol{\beta}) = \exp(\mathbf{x}_i'\boldsymbol{\beta})$, so that the cure rate for subject i is $\exp(-\theta_i) = \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))$, $i = 1, 2, \ldots, n$. This relationship between θ_i and $\boldsymbol{\beta}$ is equivalent to a canonical link for θ_i in the setting of Poisson regression models. With this relation we can write the complete data likelihood of $(\boldsymbol{\beta}, \boldsymbol{\psi})$ as

$$L(\boldsymbol{\beta}, \boldsymbol{\psi}|D) = \left(\prod_{i=1}^{n} \overline{F}(y_i|\boldsymbol{\psi})^{N_i - \nu_i}(N_i f(y_i|\boldsymbol{\psi}))^{\nu_i}\right) \times \exp\left\{\sum_{i=1}^{n}\left[N_i \mathbf{x}_i'\boldsymbol{\beta} - \log(N_i!) - \exp(\mathbf{x}_i'\boldsymbol{\beta})\right]\right\}, \quad (4.2)$$

where $D = (n, \mathbf{y}, \mathbf{X}, \boldsymbol{\nu}, \mathbf{N})$, $\mathbf{X}$ is the $n \times p$ matrix of covariates, $f(y_i|\boldsymbol{\psi})$ is the Weibull density given previously, and $\overline{F}(y_i|\boldsymbol{\psi}) = \exp(-y_i^{\alpha}\exp(\lambda))$.

Chen et al. (1999) discussed classes of noninformative priors as well as the class of power priors for $(\boldsymbol{\beta}, \boldsymbol{\psi})$ and examined some of their properties. Consider the joint noninformative prior $\pi(\boldsymbol{\beta}, \boldsymbol{\psi}) \propto \pi(\boldsymbol{\psi})$, where $\boldsymbol{\psi} = (\alpha, \lambda)'$ are the Weibull parameters in $f(y|\boldsymbol{\psi})$. This noninformative prior implies that $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ are independent a priori and $\pi(\boldsymbol{\beta}) \propto 1$ is a uniform improper prior. We will assume throughout this section that $\pi(\boldsymbol{\psi}) = \pi(\alpha|\delta_0, \tau_0)\pi(\lambda)$, where $\pi(\alpha|\delta_0, \tau_0) \propto \alpha^{\delta_0 - 1}\exp(-\tau_0\alpha)$ and δ_0 and τ_0 are two specified hyperparameters. Although several choices can be made, we will use a normal density for $\pi(\lambda)$. With these specifications the posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{\psi})$ based on the observed data $D_{\text{obs}} = (n, \mathbf{y}, \mathbf{X}, \boldsymbol{\nu})$ is given by

$$\pi(\boldsymbol{\beta}, \boldsymbol{\psi}|D_{\text{obs}}) \propto \left(\sum_{\mathbf{N}} L(\boldsymbol{\beta}, \boldsymbol{\psi}|D)\right)\pi(\alpha|\delta_0, \tau_0)\pi(\lambda), \quad (4.3)$$

where the sum in (4.3) extends over all possible values of the vector $\mathbf{N}$. Chen et al. (1999) gave conditions concerning the propriety of the posterior distribution in (4.3) using the noninformative prior $\pi(\boldsymbol{\beta}, \boldsymbol{\psi}) \propto \pi(\boldsymbol{\psi})$. This enables us to carry out Bayesian inference with improper priors for the regression coefficients and facilitates comparisons with maximum likelihood. However, under the improper priors $\pi(\boldsymbol{\beta}, \boldsymbol{\psi}) \propto \pi(\boldsymbol{\psi})$, the mixture model in (1.2) always leads to an improper posterior distribution for $\boldsymbol{\beta}$ as shown in Chen et al. (1999).

Ibrahim and Chen (2000) described a general class of informative priors called the *power priors* for conducting Bayesian inference in the presence of historical data. We now examine the power priors for $(\boldsymbol{\beta}, \boldsymbol{\psi})$. Let n_0 denote the sample size for the historical data, let $\mathbf{y}_0$ be an $n_0 \times 1$ of right-censored failure times for the historical data with censoring indicators $\boldsymbol{\nu}_0$, let $\mathbf{N}_0$ be the unobserved vector of latent counts of clonogenic cells, and let $\mathbf{X}_0$ be an $n_0 \times p$ matrix of covariates corresponding to $\mathbf{y}_0$. Let $D_0 = (n_0, \mathbf{y}_0, \mathbf{X}_0, \boldsymbol{\nu}_0, \mathbf{N}_0)$ denote the complete historical data. Further, let $\pi_0(\boldsymbol{\beta}, \boldsymbol{\psi})$ denote the initial prior distribution for $(\boldsymbol{\beta}, \boldsymbol{\psi})$. A beta prior is chosen for a_0, leading to the joint power prior distribution

$$\pi(\boldsymbol{\beta}, \boldsymbol{\psi}, a_0|D_{0,\text{obs}}) \propto \left[\sum_{\mathbf{N}_0} L(\boldsymbol{\beta}, \boldsymbol{\psi}|D_0)\right]^{a_0}\pi_0(\boldsymbol{\beta}, \boldsymbol{\psi})a_0^{\gamma_0 - 1}(1 - a_0)^{\lambda_0 - 1}, \quad (4.4)$$

where $L(\boldsymbol{\beta}, \boldsymbol{\psi}|D_0)$ is the complete data likelihood given in (4.2) with D being replaced by the historical data D_0 and $D_{0,\text{obs}} = (n_0, \mathbf{y}_0, \mathbf{X}_0, \boldsymbol{\nu}_0)$. We take a noninformative prior for $\pi_0(\boldsymbol{\beta}, \boldsymbol{\psi})$, such as $\pi_0(\boldsymbol{\beta}, \boldsymbol{\psi}) \propto \pi_0(\boldsymbol{\psi})$, which implies $\pi_0(\boldsymbol{\beta}) \propto 1$. For $\boldsymbol{\psi} = (\alpha, \lambda)'$ we take a gamma prior for α with small shape and scale parameters and an independent normal prior for λ with mean 0 and variance c_0. Also, (γ_0, λ_0) are specified prior parameters. The prior in (4.4) does not have a closed form but has several attractive properties. First, we note that if $\pi_0(\boldsymbol{\beta}, \boldsymbol{\psi})$ is proper, then (4.4) is guaranteed to be proper. Further, (4.4) can be proper even if $\pi_0(\boldsymbol{\beta}, \boldsymbol{\psi})$ is improper. Chen et al. (1999) characterized the propriety of (4.4) when $\pi_0(\boldsymbol{\beta}, \boldsymbol{\psi})$ is improper. In addition, they showed that the power prior for $\boldsymbol{\beta}$ based on the model (1.2) will always lead to an improper prior as well as an improper posterior distribution.

4.2 Semiparametric Cure Rate Model

In this section we consider a semiparametric version of the parametric cure rate model discussed in the previous section. Following Ibrahim et al. (2001a), Chen, Harrington, and Ibrahim (2001), and Chen and Ibrahim (2001), we construct a finite partition of the time axis, $0 < s_1 < \cdots < s_J$, with $s_J > y_i$ for all $i = 1, 2, \ldots, n$. Thus, we have the J intervals $(0, s_1], (s_1, s_2], \ldots, (s_{J-1}, s_J]$, and we assume that the hazard for $F(y)$ is equal to λ_j for the jth interval, $j = 1, 2, \ldots, J$, leading to

$$F^*(y|\boldsymbol{\lambda}) = 1 - \exp\left\{-\lambda_j(y - s_{j-1}) - \sum_{g=1}^{j-1}\lambda_g(s_g - s_{g-1})\right\}, \quad (4.5)$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)'$. We note that, when $J = 1$, $F^*(y|\lambda_1)$ reduces to the parametric exponential model. With this assumption the complete data likelihood can be written as

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D) &= \prod_{i=1}^{n}\prod_{j=1}^{J}\exp\left\{-(N_i - \nu_i)\nu_{ij}\left[\lambda_j(y_i - s_{j-1}) + \sum_{g=1}^{j-1}\lambda_g(s_g - s_{g-1})\right]\right\} \\ &\quad \times \prod_{i=1}^{n}\prod_{j=1}^{J}(N_i\lambda_j)^{\nu_{ij}\nu_i}\exp\left\{-\nu_i\nu_{ij}\left[\lambda_j(y_i - s_{j-1}) + \sum_{g=1}^{j-1}\lambda_g(s_g - s_{g-1})\right]\right\} \\ &\quad \times \exp\left\{\sum_{i=1}^{n}\left[N_i\mathbf{x}_i'\boldsymbol{\beta} - \log(N_i!) - \exp(\mathbf{x}_i'\boldsymbol{\beta})\right]\right\}, \end{aligned} \quad (4.6)$$

where $\nu_{ij} = 1$ if the ith subject failed or was censored in the jth interval, and 0 otherwise. The model in (4.6) is a semiparametric version of (4.2) in which the degree of the nonparametricity is controlled by J. Because the estimation of the cure rate parameter θ could be highly affected by the nonparametric nature of $F^*(y|\boldsymbol{\lambda})$, it may be desirable to choose small to moderate values of J for cure rate modeling. In practice, we recommend doing analyses for several values of J to see the sensitivity of the posterior estimates of the regression coefficients. The semiparametric cure rate model (4.6) is quite flexible, as it allows us to model general shapes of the hazard function, as well as choose the degree of parametricity in $F^*(y|\boldsymbol{\lambda})$ through suitable choices of J. Again, because $\mathbf{N}$ is not observed, the observed data likelihood, $L(\boldsymbol{\beta},\boldsymbol{\lambda}|D_{\text{obs}})$, is obtained by summing out $\mathbf{N}$ from (4.6) as in the previous section. Also, power priors for this model can be constructed in a similar fashion as in the previous section. We refer the reader to Ibrahim et al. (2001a), Chen et al. (2001), and Chen and Ibrahim (2001) for more details.

4.3 An Alternative Semiparametric Cure Rate Model

A crucial issue with cure rate modeling, and semiparametric survival models in general, is the behavior of the model in the right tail of the survival distribution. In these models there are typically few subjects at risk in the tail of the survival curve after sufficient follow-up; therefore, estimation of the cure rate can be quite sensitive to the choice of the semiparametric model. Thus, there is a need to carefully model the right tail of the survival curve and allow the model to be more parametric in the tail, while also allowing the model to be nonparametric in other parts of the curve. Ibrahim, Chen, and Sinha (2001b) constructed such a model by defining a smoothing parameter κ, $0<\kappa<1$, that controls the degree of parametricity in the right tail of the survival curve and does not depend on the data. Specifically, the prior for λ_j depends on κ, such that the model converges to a parametric model in the right tail of the survival curve as $j\to\infty$. By an appropriate choice of κ, one can choose a fully nonparametric model or a fully parametric model for the right tail of the survival distribution. Also, κ will allow some control over the degree of parametricity in the beginning and middle part of the survival distribution. A more parametric shape of the model in the right tail facilitates more stable and precise estimates of the parameters. This approach is fundamentally very different from previous approaches for semiparametric Bayesian survival analysis, which primarily focus on specifying a prior process with a mean function and possibly a smoothing parameter, in which the posterior properties of both of them depend on the data.

Let $F_0(t|\boldsymbol{\psi}_0)$ denote the parametric survival model chosen for the right tail of the survival curve and let $H_0(t)$ denote the corresponding cumulative baseline hazard function. Now take the λ_j's to be independent a priori, each having a gamma prior distribution with mean

$$\mu_j=E(\lambda_j|\boldsymbol{\psi}_0)=\frac{H_0(s_j)-H_0(s_{j-1})}{s_j-s_{j-1}}, \qquad (4.7)$$

and variance

$$\sigma_j^2=\operatorname{Var}(\lambda_j|\boldsymbol{\psi}_0,\kappa)=\mu_j\kappa^j, \qquad (4.8)$$

where $0<\kappa<1$ is the smoothing parameter. As $\kappa\to0$, $\sigma_j^2\to0$, so that small values of κ imply a more parametric model in the right tail. In addition, as $j\to\infty$, $\sigma_j^2\to0$, implying that the degree of parametricity is increased at a rate governed by κ as the number of intervals increases. This property also implies that, as $j\to\infty$, the survival distribution in the right tail becomes more parametric regardless of any fixed value of κ. The properties of this model are attractive. For example, if $F_0(\cdot|\boldsymbol{\psi}_0)$ is an exponential distribution, then $F_0(y|\psi_0)=1-\exp(-\psi_0y)$, so that $\mu_j=\psi_0$ and $\sigma_j^2=\psi_0\kappa^j$. If $F_0(\cdot|\boldsymbol{\psi}_0)$ is a Weibull distribution, then $F_0(y|\boldsymbol{\psi}_0)=1-\exp(-\gamma_0y^{\alpha_0})$, $\boldsymbol{\psi}_0=(\alpha_0,\gamma_0)'$, so that

$$\mu_j=\gamma_0\frac{(s_j^{\alpha_0}-s_{j-1}^{\alpha_0})}{s_j-s_{j-1}} \quad\text{and}\quad \sigma_j^2=\gamma_0\frac{(s_j^{\alpha_0}-s_{j-1}^{\alpha_0})}{s_j-s_{j-1}}\kappa^j.$$

Several properties of this model are now given, which are proved in Ibrahim et al. (2001a,b).

Property 4.1. Assume that $(s_j+s_{j-1})/2\to t$ as $s_j-s_{j-1}\to0$. Then, for any j, according to this prior process, $E(\lambda_j|\boldsymbol{\psi}_0)\to h_0(t)$ as $s_j-s_{j-1}\to0$, where $h_0(t)=\frac{d}{dt}H_0(t)$.

For example, if $F_0(y|\psi_0)=1-\exp(-\psi_0y)$, then $E(\lambda_j|\psi_0)=\psi_0$ regardless of the choice of $s_1,s_2,\ldots,s_J$. If $F_0(y|\boldsymbol{\psi}_0)=1-\exp(-\gamma_0y^{\alpha_0})$, then $E(\lambda_j|\boldsymbol{\psi}_0)\to\gamma_0\alpha_0t^{\alpha_0-1}$ as $s_j-s_{j-1}\to0$. This assures that, as j becomes large and $s_j-s_{j-1}\to0$, this prior process approximates any prior process with prior mean $h_0(t)$ defined on the progression time hazard $h^*(t|\boldsymbol{\lambda})$ corresponding to (4.5).

Property 4.2. Let $\overline{F}_p^*(y|\boldsymbol{\lambda})=\exp(-\theta F^*(y|\boldsymbol{\lambda}))$. Then $\overline{F}_p^*(y|\boldsymbol{\lambda})\to\overline{F}_p(y|\boldsymbol{\psi}_0)$ as $\kappa\to0$, where $\overline{F}_p(y|\boldsymbol{\psi}_0)=\exp(-\theta F_0(y|\boldsymbol{\psi}_0))$.

Property 4.3. Let $f^*(y|\boldsymbol{\lambda})=\frac{d}{dy}F^*(y|\boldsymbol{\lambda})$ and let $h_p^*(y|\boldsymbol{\lambda})=\theta f^*(y|\boldsymbol{\lambda})$ denote the corresponding hazard function. Then $h_p^*(y|\boldsymbol{\lambda})\to\theta f_0(y|\boldsymbol{\psi}_0)$ as $\kappa\to0$, where $f_0(y|\boldsymbol{\psi}_0)=\frac{d}{dy}F_0(y|\boldsymbol{\psi}_0)$.

We now give joint prior specifications for the semiparametric model in (4.7) and (4.8). We specify a hierarchical model and first consider a joint (improper) noninformative prior distribution for $(\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\psi}_0)$, given by

$$\pi(\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\psi}_0)=\pi(\boldsymbol{\beta})\pi(\boldsymbol{\lambda}|\boldsymbol{\psi}_0)\pi(\boldsymbol{\psi}_0)$$
$$\propto\pi(\boldsymbol{\beta})\left[\prod_{j=1}^J\pi(\lambda_j|\boldsymbol{\psi}_0)\right]\pi(\boldsymbol{\psi}_0). \qquad (4.9)$$

We take each $\pi(\lambda_j|\boldsymbol{\psi}_0)$ to be independent gamma densities with mean μ_j and variance σ_j^2. If $F_0(\cdot|\psi_0)$ is an exponential distribution, then ψ_0 is a scalar, and we specify a gamma prior for it; that is, $\pi(\psi_0)\propto\psi_0^{\zeta_0-1}\exp(-\tau_0\psi_0)$, where ζ_0 and τ_0 are specified hyperparameters. If $F_0(\cdot|\boldsymbol{\psi}_0)$ is a Weibull distribution, then $\boldsymbol{\psi}_0=(\alpha_0,\gamma_0)'$. In this case we take a prior of the form

$$\pi(\boldsymbol{\psi}_0)=\pi(\alpha_0,\gamma_0)$$
$$\propto\alpha_0^{\zeta_{\alpha_0}-1}\exp(-\tau_{\alpha_0}\alpha_0)\gamma_0^{\zeta_{\gamma_0}-1}\exp(-\tau_{\gamma_0}\gamma_0), \qquad (4.10)$$

where ζ_{α_0}, τ_{α_0}, ζ_{γ_0}, and τ_{γ_0} are specified hyperparameters. For $\boldsymbol{\beta}$ we consider a uniform improper prior. Ibrahim et al. (2001a, 2001b) established conditions for the propriety of the joint posterior distribution of $(\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\psi}_0)$, when using an exponential distribution or a Weibull distribution for $F_0(\cdot|\boldsymbol{\psi}_0)$. In addition, they developed the power prior for this model.

4.4 Multivariate Cure Rate Model

It is often of interest to jointly model several types of failure time random variables in survival analysis, such as time to cancer relapse at two different organs, times to cancer relapse and death, times to first and second infection, and so forth. In addition, these random variables typically have joint and marginal survival curves that "plateau" beyond a certain period of follow-up; therefore, it is of great importance in these situations to develop a joint cure rate model for inference. There does not appear to be a natural multivariate extension of the mixture model in (1.2). Even if such an extension were available, it appears that a multivariate mixture model would be extremely cumbersome to work with from a theoretical and computational perspective. As an alternative to a direct multivariate extension of (1.2), Chen, Ibrahim, and Sinha (2001) proposed a multivariate generalization of (4.1), called the *multivariate cure rate model.* This model proves to be quite useful for modeling multivariate data in which the joint failure time random variables have a surviving fraction and each marginal failure time random variable also has a surviving fraction. To induce the correlation structure between the failure times, Chen et al. (2001) introduced a frailty term (Clayton 1978; Hougaard 1986; Oakes 1989), which is assumed to have a positive stable distribution. A positive stable frailty results in a conditional proportional hazards structure (i.e., given the unobserved frailty). Thus, the marginal and conditional hazards of each component have a proportional hazards structure and remain in the same class of univariate cure rate models.

For clarity and ease of exposition, we will focus our discussion on the bivariate cure rate model, as extensions to the general multivariate case are quite straightforward. The bivariate cure rate model of Chen et al. (2001) can be derived as follows. Let $\mathbf{Y} = (Y_1, Y_2)'$ be a bivariate failure time, such as Y_1 = time to cancer relapse and Y_2 = time to death, or Y_1 = time to first infection and Y_2 = time to second infection, and so forth. We assume that (Y_1, Y_2) are not ordered and have support on the upper orthant of the plane. For an arbitrary patient in the population, let $\mathbf{N} = (N_1, N_2)'$ denote latent (unobserved) variables for (Y_1, Y_2), respectively. We assume throughout that N_k has a Poisson distribution with mean $\theta_k w$, $k = 1, 2$, and (N_1, N_2) are independent, given w. The quantity w is a frailty component in the model that induces a correlation between the latent variables (N_1, N_2). Here we take w to have a positive stable distribution indexed by the parameter α, denoted by $w \sim S_\alpha(1, 1, 0)$, where $0 < \alpha < 1$ (see Chen et al. 2001 for more details). Although several choices can be made for the distribution of w, the positive stable distribution is quite attractive, common, and flexible in the multivariate survival setting. In addition, it will yield several desirable properties.

Let $\mathbf{Z}_i = (Z_{1i}, Z_{2i})'$ denote the bivariate progression time for the ith clonogenic cell. The random vectors $\mathbf{Z}_i$, $i = 1, 2, \ldots$, are assumed to be independent and identically distributed. The cumulative distribution function of Z_{ki} is denoted by $F_k(t) = 1 - \overline{F}_k(t)$, $k = 1, 2$, and F_k is independent of (N_1, N_2). The observed survival time can be defined by the random variable $Y_k = \min\{Z_{ki}, 0 \le i \le N_k\}$, where $P(Z_{k0} = \infty) = 1$ and N_k is independent of the sequence $Z_{k1}, Z_{k2}, \ldots$, for $k = 1, 2$. The survival function for $\mathbf{Y} = (Y_1, Y_2)'$, given w, and hence the survival function for the population, given w, is given by

$$\begin{aligned}
&\overline{F}_{\text{pop}}(y_1, y_2|w)\\
&\quad= \prod_{k=1}^{2} \left[P(N_k = 0) + P(Z_{k1} > y_k, \ldots, Z_{kN} > y_k, N_k \ge 1)\right]\\
&\quad= \prod_{k=1}^{2} \left[\exp(-w\theta_k) + \left(\sum_{r=1}^{\infty} \overline{F}_k(y_k)^r \, \frac{(w\theta_k)^r}{r!} \exp(-w\theta_k)\right)\right]\\
&\quad= \prod_{k=1}^{2} \exp\{-w\theta_k + \theta_k w \overline{F}_k(y_k)\}\\
&\quad= \exp\{-w[\theta_1 F_1(y_1) + \theta_2 F_2(y_2)]\}, \qquad (4.11)
\end{aligned}$$

where $P(N_k = 0) = P(Y_k = \infty) = \exp(-\theta_k)$, $k = 1, 2$. The frailty variable w serves a dual purpose in the model—it induces the correlation between Y_1 and Y_2 and at the same time relaxes the Poisson assumption of N_1 and N_2 by adding the same extra Poisson variation through their respective means $\theta_1 w$ and $\theta_2 w$.

The Laplace transform of w is given by $E(\exp(-sw)) = \exp(-s^\alpha)$. Using the Laplace transform of w, a straightforward derivation yields the unconditional survival function

$$\overline{F}_{\text{pop}}(y_1, y_2) = \exp\{-[\theta_1 F_1(y_1) + \theta_2 F_2(y_2)]^\alpha\}. \qquad (4.12)$$

It can be shown that (4.12) has a proportional hazards structure if the covariates enter the model through (θ_1, θ_2). The joint cure fraction implied by (4.12) is $\overline{F}_{\text{pop}}(\infty, \infty) = \exp(-[\theta_1 + \theta_2]^\alpha)$. From (4.12) the marginal survival functions are $\overline{F}_k(y) = \exp(-\theta_k^\alpha (F_k(y))^\alpha)$, $k = 1, 2$. Equation (4.12) indicates that the marginal survival functions have a cure rate structure with probability of cure $\exp(-\theta_k^\alpha)$ for Y_k, $k = 1, 2$. It is important to note in (4.12) that each marginal survival function has a proportional hazards structure as long as the covariates, $\mathbf{x}$, only enter through θ_k. The marginal hazard function is given by $\alpha\theta_k^\alpha f_k(y)(F_k(y))^{\alpha-1}$, with attenuated covariate effect $(\theta_k(x))^\alpha$, and $f_k(y)$ is the survival density corresponding to $F_k(y)$. This property is similar to the earlier observations made by Oakes (1989) for the ordinary bivariate stable frailty survival model. The parameter α, $0 < \alpha < 1$, is a scalar parameter that is a measure of association between (Y_1, Y_2). Small values of α indicate high association between (Y_1, Y_2). As $\alpha \to 1$ this implies less association between (Y_1, Y_2), which can be seen from (4.12). Following Clayton (1978) and Oakes (1989), we can compute a local measure of dependence, denoted $\theta^*(y_1, y_2)$, as a function of α. For the multivariate cure rate model in (4.12), $\theta^*(t_1, t_2)$ is well defined and is given by

$$\theta^*(y_1, y_2) = \alpha^{-1}(1 - \alpha)\big(\theta_1 F_1(y_1) + \theta_2 F_2(y_2)\big)^{-\alpha} + 1. \quad (4.13)$$

We see that $\theta^*(y_1, y_2)$ in (4.13) decreases in (y_1, y_2). That is, the association between (Y_1, Y_2) is greater when (Y_1, Y_2) are small and the association decreases over time. Such a property is desirable, for example, when Y_1 denotes time to relapse and Y_2 denotes time to death. Finally, we mention that a global measure of dependence such as Kendall's τ or the Pearson correlation coefficient is not well defined for the multivariate cure rate model (4.12) because no moments for cure rate models exist due to the improper survival function.

The likelihood function for this model based on n subjects can be written as

$$\begin{aligned}
&L(\boldsymbol{\theta}, \boldsymbol{\psi}|D)\\
&\quad= \left(\prod_{k=1}^{2}\prod_{i=1}^{n} \overline{F}_k(y_{ki}|\boldsymbol{\psi}_k)^{N_{ki}-\nu_{ki}} \big(N_{ki} f_k(y_{ki}|\boldsymbol{\psi}_k)\big)^{\nu_{ki}}\right)\\
&\qquad \times \exp\left\{\sum_{i=1}^{n} \big(N_{ki}\log(w_i\theta_k) - \log(N_{ki}!) - w_i\theta_k\big)\right\}, \quad (4.14)
\end{aligned}$$

where $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)$, $\boldsymbol{\theta} = (\theta_1, \theta_2)$, N_{ki} is the number of clonogens for the ith subject, and $f_k(y_{ki}|\boldsymbol{\psi}_k)$ is the density corresponding to $F_k(y_{ki}|\boldsymbol{\psi}_k)$, $i = 1, \ldots, n$, $k = 1, 2$.

Chen et al. (2001) used a Weibull density for $f_k(y_{ki}|\boldsymbol{\psi}_k)$, so that $f_k(y|\boldsymbol{\psi}_k) = \xi_k y^{\xi_k - 1} \exp\{\lambda_k - y^{\xi_k} \exp(\lambda_k)\}$, where $\boldsymbol{\psi}_k = (\xi_k, \lambda_k)$, $k = 1, 2$. To construct the likelihood function of the observed data, we integrate (4.14) with respect to $(\mathbf{N}, \mathbf{w})$ assuming an $S_\alpha(1, 1, 0)$ density for each w_i, denoted by $f_s(w_i|\alpha)$, where $\mathbf{N} = (N_{11}, \ldots, N_{1n}, N_{21}, \ldots, N_{2n})$ and $\mathbf{w} = (w_1, \ldots, w_n)$. As before we incorporate covariates for the cure rate model (4.12) through the cure rate parameter $\boldsymbol{\theta}$. Let $\mathbf{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ denote the $p \times 1$ vector of covariates for the ith subject and let $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kp})'$ denote the corresponding vector of regression coefficients for failure time random variable Y_k, $k = 1, 2$. We relate $\boldsymbol{\theta}$ to the covariates by $\theta_{ki} \equiv \theta(\mathbf{x}_i'\boldsymbol{\beta}_k) = \exp(\mathbf{x}_i'\boldsymbol{\beta}_k)$, so that the cure rate for subject i is $\exp(-\theta_{ki}) = \exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}_k))$ for $i = 1, 2, \ldots, n$ and $k = 1, 2$. Letting $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, Chen et al. (2001) showed that the observed data likelihood of $(\boldsymbol{\beta}, \boldsymbol{\psi}, \alpha)$ can be written as

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\psi}, \alpha | D_{\text{obs}}) &= \left(\alpha^{d_1+d_2} \prod_{k=1}^{2} \prod_{i \in \mathcal{D}_k} \exp(\mathbf{x}_i'\boldsymbol{\beta}_k)\right) \left[\prod_{k=1}^{2} \prod_{i=1}^{n} f_k(y_{ki}|\boldsymbol{\psi}_k)^{\nu_{ki}}\right] \\ &\times \prod_{i=1}^{n} \Big\{ \big[\exp(\mathbf{x}_i'\boldsymbol{\beta}_1) F_1(y_{1i}|\boldsymbol{\psi}_1) \\ &\qquad + \exp(\mathbf{x}_i'\boldsymbol{\beta}_2) F_2(y_{2i}|\boldsymbol{\psi}_2)\big]^{(\alpha-1)(\nu_{1i}+\nu_{2i})} \Big\} \\ &\times \prod_{i=1}^{n} \Big\{ \frac{1-\alpha}{\alpha} \big[\exp(\mathbf{x}_i'\boldsymbol{\beta}_1) F_1(y_{1i}|\boldsymbol{\psi}_1) \\ &\qquad + \exp(\mathbf{x}_i'\boldsymbol{\beta}_2) F_2(y_{2i}|\boldsymbol{\psi}_2)\big]^{-\alpha} + 1 \Big\}^{\nu_{1i}\nu_{2i}} \\ &\times \prod_{i=1}^{n} \exp\Big\{ -\big(\exp(\mathbf{x}_i'\boldsymbol{\beta}_1) F_1(y_{1i}|\boldsymbol{\psi}_1) \\ &\qquad + \exp(\mathbf{x}_i'\boldsymbol{\beta}_2) F_2(y_{2i}|\boldsymbol{\psi}_2)\big)^{\alpha} \Big\}, \qquad (4.15) \end{aligned}$$

where $\mathcal{D}_k$ consists of those patients who failed according to Y_k, $k = 1, 2$, $D_{\text{obs}} = (n, \mathbf{y}_1, \mathbf{y}_2, \mathbf{X}, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2)$, $\mathbf{X}$ is the $n \times p$ matrix of covariates, $f_k(y_{ki}|\boldsymbol{\psi}_k)$ is Weibull, and $\overline{F}_k(y_{ki}|\boldsymbol{\psi}_k) = \exp(-y_{ki}^{\xi_k} \exp(\lambda_k))$.

Chen et al. (2001) considered a joint improper prior for $(\boldsymbol{\beta}, \boldsymbol{\psi}, \alpha) = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \alpha)$ of the form

$$\begin{aligned} \pi(\boldsymbol{\beta}, \boldsymbol{\psi}, \alpha) &= \pi(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \alpha) \\ &\propto \pi(\boldsymbol{\psi}_1)\pi(\boldsymbol{\psi}_2) I(0 < \alpha < 1) \\ &= \prod_{k=1}^{2} \pi(\xi_k, \lambda_k) I(0 < \alpha < 1), \end{aligned}$$

where $I(0 < \alpha < 1) = 1$ if $0 < \alpha < 1$, and 0 otherwise. This prior implies that $\boldsymbol{\beta}$, $\boldsymbol{\psi}$, and α are independenta priori, $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ are independent a priori with an improper uniform prior, α has a proper uniform prior over the interval $(0, 1)$, and $(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)$ are independent and identically distributed as $\pi(\boldsymbol{\psi}_k)$ a priori. Chen et al. (2001) assumed that $\pi(\xi_k, \lambda_k) = \pi(\xi_k|\nu_0, \tau_0)\pi(\lambda_k)$, where

$$\pi(\xi_k|\delta_0, \tau_0) \propto \xi_k^{\delta_0 - 1} \exp\{-\tau_0 \xi_k\} \quad \text{and} \quad \pi(\lambda_k) \propto \exp\{-c_0 \lambda_k^2\},$$

and δ_0, τ_0, and c_0 are specified hyperparameters. Chen et al. (2001) and Ibrahim et al. (2001a) gave a detailed discussion of the computational implementation of this model.

Detailed real data examples for the models discussed in Sections 4.1–4.3 can be found in Ibrahim et al. (2001a). The Gibbs sampler was used to obtained posterior estimates for all models discussed in this section. Details of the Gibbs sampling algorithms and covergence diagnostics can be found in Ibrahim et al. (2001).

5. FUTURE RESEARCH

Future areas of research include expanding the Bayesian model to include covariates in θ as well as $F(t)$. Although not investigated here, this appears to be a natural extension of the Bayesian model presented in Section 4. A more general regression model is currently being investigated. The results reported in Section 3 from the frequentist perspective also encourage this research effort.

The class of models given by expression (1.8) in Section 1 opens a new avenue in parametric (and probably semiparametric; see Section 3) survival regression. For example, by incorporating a predictor into the parameter i we obtain a proportional hazards (PH) model. The quantity i is unobservable, but it is reasonable to assume that i is proportional to the observable tumor volume. If one predictor is incorporated into the parameter i whereas another covariate (e.g., fractional radiation dose) is incorporated into some other parameter(s) of the model, the resulting regression counterpart of (1.8) will no longer be the PH model. From a statistical viewpoint the main advantage of this generalization of the clonal model of tumor recurrence is that it can suggest the structure of new regression models to be further explored by statistical methods. Such models may provide a useful means of searching for better regimens for fractionated radiotherapy.

Another promising idea is to explore possible links between stochastic models proposed for the natural history of cancer and for posttreatment cancer survival; this approach can enrich statistical inference by supplementing the analysis of patients' survival with additional epidemiological data on cancer detection. We believe that this idea may dramatically change the whole concept of modern parametric survival analysis, by invoking mechanistically motivated models for the joint distribution of covariates at the time of diagnosis.

Methods for model diagnostics especially designed for the semiparametric models in Section 3 remain a very important issue for future methodological research. The same is equally true for the regression counterparts of the two-component mixture model. Justification of asymptotic properties of the estimate represent a serious challenge.

6. SOFTWARE AND DATA ANALYSIS FOR CURE RATE MODELS

Prototype software implementing the estimation procedures of Section 3.2 was programmed in Delphi 6.0 for Windows 95/98/2000/XP, an object-oriented visual development environment based on the Pascal programming language by INPRISE™. This software is available on request from Alex Tsodikov. Implementation of these methods in high-level languages such as R, S, or SAS appears straightforward. In addition, Extensive BUGS software is available for the cure rate and other related models given in the book by Ibrahim et al. (2001a). This software can be easily downloaded from the book's Web site at *http://www.merlot.uconn.edu/mhchen/survbook.* BUGS code for several models and datasets is available from this Web site.

In the semiparametric context a cure model can be regarded as a convenient reparameterization of its counterpart in the noncure form [compare the traditional form of the PH model $G = \overline{F}^{\theta}$ with its cure form (3.1)]. A display of the Kaplan–Meier estimator for the data may indicate a noncure situation if the curves go to 0 at the end of follow-up. Fitting a cure form model to such data may result in numerical instability as the parameter coding the cure rate would be approaching the border of the parametric space.

Departures from proportionality and an indication for the extended hazard models can be explored using an empirical display of the function F (2.4) or its more sophisticated versions as discussed in Section 2. Similar techniques for model exploration is available for two-component mixture models (Tsodikov 2002). More work is needed to create a toolbox of methods for model building, choice and diagnostics.

7. BIBLIOGRAPHICAL NOTES

Over recent years the BCH model has been developed in many different directions. In particular, the literature on parametric inference based on this model is quite voluminous. The usual practice is to use the two-parameter gamma or the Weibull distribution for the function $F(t)$ in formula (1.4). Maximum likelihood inference without covariates has been discussed in the context of right-censored data under continuous follow-up (Asselain, Fourquet, Hoang, Myasnikova, and Yakovlev 1994; Hoang, Tsodikov, Yakovlev, and Asselain 1995; Yakovlev 1996; Yakovlev and Tsodikov 1996); discrete surveillance design (Tsodikov, Asselain, Fourquet, Hoang, and Yakovlev 1995), and doubly censored data (Kruglikov, Pilipenko, Tsodikov, and Yakovlev 1997). A version of the Hjort goodness-of-fit test for the model (1.4) with $F(t)$ represented by the gamma distribution was described in Gregori et al. (2002). Tsodikov (1998a) studied the asymptotic efficiency of cure rate estimation. Some limiting distributions associated with model (4.1) and their bivariate counterparts were described in the articles by Klebanov, Rachev, and Yakovlev (1993a) and Rachev, Wu, and Yakovlev (1995); the authors also provided estimates of the convergence rates. Parametric and semiparametric regression models allowing for dissimilar effects of covariates on the probability of cure and the timing of the event of interest were extensively explored in several publications (Asselain, Fourquet, Hoang, Tsodikov, and Yakovlev 1996; Tsodikov et al. 1998b; Myasnikova, Asselain, and Yakovlev 2000; Tsodikov 2002). An improper proportional hazards model was studied by Tsodikov (1998a,b) and Chen and Ibrahim (2001). Time-dependent risk factors for the cure probability were introduced in Tsodikov et al. (1997, 1998b) and Tsodikov and Müller (1998). Two-sample score tests for long-term and short-term effects on survival were proposed by Broët et al. (2001). Statistical inference from the Bayesian prospective was discussed by Klebanov, Rachev, and Yakovlev (1993b), Chen et al. (1999, 2001), Ibrahim and Chen (2000), Ibrahim et al. (2001a,b), and Chen, Ibrahim, and Lipsitz, (2002). Ibrahim et al. (2001a) devoted an entire chapter (chap. 5) to the cure rate model and its applications in their book. Moreover, recent books discussing Gibbs sampling for this model and related models are also given in Ibrahim et al. (2001a) and Chen, Shao, and Ibrahim (2000). The model has successfully been applied to various datasets on cancer survival (Yakovlev et al. 1993; Asselain et al. 1994, 1996; Yakovlev 1996; Yakovlev and Tsodikov 1996; Tsodikov et al. 1995, 1997; Tsodikov 1998a, Tsodikov et al. 1998b; Chen et al. 1999; Yakovlev et al. 1999; Myasnikova et al. 2000; Trelford, Tsodikov, and Yakovlev 2001; Tsodikov 2001, 2002; Tsodikov, Dicello, Zaider, Zorin, and Yakovlev 2001; Zaider et al. 2001; Zorin, Tsodikov, Zharinov, and Yakovlev 2001).

[Received July 2002. Revised June 2003.]

REFERENCES

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1992), *Statistical Models Based on Counting Processes*, New York: Springer-Verlag.

Asselain, B., Fourquet, A., Hoang, T., Myasnikova, C., and Yakovlev, A. Yu. (1994), "Testing the Independence of Competing Risks: An Application to the Analysis of Breast Cancer Recurrence," *Biometrical Journal*, 36, 465–473.

Asselain, B., Fourquet, A., Hoang, T., Tsodikov, A., and Yakovlev, A. Yu. (1996), "A Parametric Regression Model of Tumor Recurrence: An Application to the Analysis of Clinical Data on Breast Cancer," *Statistics & Probability Letters*, 29, 271–278.

Bartoszyński, R., Edler, L., Hanin, L., Kopp-Schneider, A., Pavlova, L., Tsodikov, A., Zorin, A., and Yakovlev, A. (2001), "Modeling Cancer Detection: Tumor Size as a Source of Information on Unobservable Stages of Carcinogenesis," *Mathematical Biosciences*, 171, 113–142.

Bentzen, S. M., Johansen, L. V., Overgaard, J., and Thames, H. D. (1991), "Clinical Radiobiology of Squamous Cell Carcinoma of the Oropharynx," *International Journal Radiation. Oncology–Biology–Physics*, 20, 1197–1206.

Berkson, J., and Gage, R. P. (1952), "Survival Curves for Cancer Patients Following Treatment," *Journal of the American Statistical Association*, 47, 501–515.

Boag, J. M. (1949), "Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy," *Journal of the Royal Statistical Society*, Ser. B, 11, 15–44.

Broët, P., Rycke, Y. D., Tubert-Bitter, P., Lellouch, J., Asselain, B., and Moreau, T. (2001), "A Semiparametric Approach for the Two-Sample Comparison of Survival Times With Long-Term Survivors," *Biometrics*, 57, 844–852.

Cantor, A. B. (1997), "Cure Rate Estimation Using the Polynomial Gompertz Model," in *The Communication of Statitistics: Exchange and Dissemination*, 1997 Spring Meeting of the International Biometric Society, Eastern North American Region, Abstracts, p. 144.

——— (2001), "Projecting the Standard Error of the Kaplan–Meier Estimator," *Statistics in Medicine*, 20, 2091–2097.

Cantor, A. B., and Shuster, J. J. (1992), "Parametric Versus Nonparametric Methods for Estimating Cure Rates Based on Censored Survival Data," *Statistics in Medicine*, 11, 931–937.

Chappell, R., Nondahl, D. M., and Fowler, J. F. (1995), "Modeling Dose and Local Control in Radiotherapy," *Journal of the American Statistical Association*, 90, 829–838.

Chen, M. H., Harrington, D. P., and Ibrahim, J. G. (2002a), "Bayesian Cure Rate Models for Malignant Melanoma: A Case Study of ECOG Trial E1690," *Applied Statistics*, 51, 135–150.

Chen, M. H., and Ibrahim, J. G. (2001), "Maximum Likelihood Methods for Cure Rate Models With Missing Covariates," *Biometrics*, 57, 43–52.

Chen, M. H., Ibrahim, J. G., and Lipsitz, S. R. (2002b), "Bayesian Methods for Missing Covariates in Cure Rate Models," *Lifetime Data Analysis*, 8, 117–146.

Chen, M. H., Ibrahim, J. G., and Sinha, D. (1999), "A New Bayesian Model for Survival Data With a Surviving Fraction," *Journal of the American Statistical Association*, 94, 909–919.

——— (2002c), "Bayesian Inference for Multivariate Survival Data With a Surviving Fraction," *Journal of Multivariate Analysis*, 80, 101–126.

Chen, M. H., Shao, Q. M., and Ibrahim, J. G. (2000), *Monte Carlo Methods in Bayesian Computation*, New York: Springer-Verlag.

Cheng, S., Wei, L., and Ying Z. (1995), "Analysis of Transformation Models With Censored Data," *Biometrika*, 82, 835–845.

Clayton, D. G. (1978), "A Model for Association in Bivariate Life Tables and Its Application in Epidemiological Studies in Familial Tendency in Chronic Disease Incidence," *Biometrika*, 65, 141–151.

Cox, D. R. (1972), "Regression Models and Life Tables" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 34, 187–220.

Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Etezadi-Amoli, L., and Ciampi, A. (1987), "Extended Hazard Regression for Censored Survival Data With Covariates: A Spline Approximation for the Baseline Hazard Function," *Biometrics*, 43, 181–192.

Farewell, V. T. (1982), "The Use of Mixture Models for the Analysis of Survival Data With Long-Term Survivors," *Biometrics*, 38, 1041–1046.

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications*, Vol. 2, New York: Wiley.

Fleming, T., and Lin, D. (2000), "Survival Analysis in Clinical Trials: Past Developments and Future Directions," *Biometrics*, 56, 971–983.

Gamel, J. W., McLean, I. W., and Rosenberg, S. H. (1990), "Proportion Cured and Mean Log Survival Time as Functions of Tumour Size," *Statistics in Medicine*, 9, 999–1006.

Gamel, J. W., Meyer, J. S., Feuer, E., and Miller, B. A. (1996), "The Impact of Stage and Histology on the Long-Term Clinical Course of 163,808 Patients With Breast Carcinoma," *Cancer*, 77, 1459–1469.

Gamel, J. W., and Vogel, R. L. (1993), "A Model of Long-Term Survival Following Adjuvant Therapy for Stage 2 Breast Cancer," *British Journal of Cancer*, 68, 1167–1170.

Gamel, J. W., Vogel, R. L., and McLean, I. W. (1993), "Assessing the Impact of Adjuvant Therapy on Cure Rate for Stage 2 Breast Carcinoma," *British Journal of Cancer*, 68, 115–118.

Gamel, J. W., Vogel, R. L., Valagussa, P., and Bonadonna, G. (1994), "Parametric Survival Analysis of Adjuvant Therapy for Stage II Breast Cancer," *Cancer*, 74, 2483–2490.

Goldman, A. I. (1984), "Survivorship Analysis When Cure Is a Possibility: A Monte-Carlo Study," *Statistics in Medicine*, 3, 153–163.

Goldman, A. I., and Hillman, D. W. (1992), "Exemplary Data: Sample Size and Power in the Design of Event-Time Clinical Trials," *Controlled Clinical Trials*, 13, 256–271.

Gordon, N. H. (1990), "Application of the Theory of Finite Mixtures for the Estimation of 'Cure' Rates of Treated Cancer Patients," *Statistics in Medicine*, 9, 397–407.

Greenhouse, J. B., and Wolfe, R. A. (1984), "A Competing Risk Derivation of a Mixture Model for the Analysis of Survival Data," *Communications in Statistics—Theory and Methods*, 25, 3133–3154.

Gregori, G., Hanin, L., Luebeck, G., Moolgavkar, S., and Yakovlev, A. (2002), "Testing Goodness of Fit With Stochastic Models of Carcinogenesis," *Mathematical Biosciences*, 175, 13–29.

Hanin, L. G. (2001), "Iterated Birth and Death Process as a Model of Radiation Cell Survival," *Mathematical Biosciences*, 169, 89–107.

Hanin, L. G., Zaider, M., and Yakovlev, A. Y. (2001), "Distribution of the Number of Clonogens Surviving Fractionated Radiotherapy: A Long-Standing Problem Revisited," *International Journal of Radiation Biology*, 77, 205–213.

Haybittle, J. L. (1959), "The Estimation of the Proportion of Patients Cured After Treatment for Cancer of the Breast," *British Journal of Radiology*, 32, 725–733.

——— (1965), "A Two-Parameter Model for the Survival Curve of Treated Cancer Patients," *Journal of the American Statistical Association*, 60, 16–26.

Hjort, N. (1990), "Goodness of Fit Tests in Models for Life History Based on Cumulative Hazard Rates," *The Annals of Statistics* 18, 1221–1258.

Hoang, T., Tsodikov, A., Yakovlev, A. Yu., and Asselain, B. (1995), "Modeling Breast Cancer Recurrence," in *Mathematical Population Dynamics: Analysis of Heterogeneity*, Vol. 2, eds. O. Arino, D. Axelrod, and M. Kimmel, Winnipeg: Wuerz Publications, pp. 283–296.

Hougaard, P. (1986), "A Class of Multivariate Failure Time Distributions," *Biometrika*, 73, 671–678.

Ibrahim, J. G., and Chen, M. H. (2000), "Power Prior Distributions for Regression Models," *Statistical Science*, 15, 46–60.

Ibrahim, J. G., Chen, M. H., and Sinha, D. (2001a), *Bayesian Survival Analysis*, New York: Springer-Verlag.

——— (2001b), "Bayesian Semi-Parametric Models for Survival Data With a Cure Fraction," *Biometrics*, 57, 383–388.

Kalbfleisch, J. D., and Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data*, New York: Wiley.

Klebanov, L. B., Rachev, S. T., and Yakovlev, A. Yu. (1993a), "A Stochastic Model of Radiation Carcinogenesis: Latent Time Distributions and Their Properties," *Mathematical Biosciences*, 113, 51–75.

——— (1993b), "On the Parametric Estimation of Survival Functions," *Statistics & Decisions*, 3, 83–102.

Klein, J. (1992), "Semiparametric Estimation of Random Effects Using the Cox Model Based on the EM Algorithm," *Biometrics*, 48, 795–806.

Kruglikov, I. L., Pilipenko, N. I., Tsodikov, A. D., and Yakovlev, A. Y. (1997), "Assessing Risk With Doubly Censored Data: An Application to the Analysis of Radiation-Induced Thyropathy," *Statistics & Probability Letters*, 32, 223–230.

Kuk, A. Y. C., and Chen, C.-H. (1992), "A Mixture Model Combining Logistic Regression With Proportional Hazards Regression," *Biometrika*, 79, 531–541.

Lange, K., Hunter, D. R., and Yang, I. (2000), "Optimization Transfer Using Surrogate Objective Functions" (with discussion), *Journal of Computational and Graphical Statistics*, 9, 1–59.

Laska, E. M., and Meisner, M. J. (1992), "Nonparametric Estimation and Testing in a Cure Model," *Biometrics*, 48, 1223–1234.

Maller, R. A., and Zhou, S. (1992), "Estimating the Proportion of Immunes in a Censored Sample," *Biometrika*, 79, 731–739.

——— (1994), "Testing for Sufficient Follow-up and Outliers in Survival Data," *Journal of the American Statistical Association*, 89, 1499–1506.

——— (1995), "Testing for the Presence of Immune or Cured Individuals in Censored Survival Data," *Biometrics*, 51, 1197–1205.

——— (1996), *Survival Analysis With Long-Term Survivors*, Chichester, U.K.: Wiley.

Miller, R. G. (1981), *Survival Analysis*, New York: Wiley.

Murphy, S. (2000), "On Profile Likelihood," *Journal of the American Statistical Association*, 95, 449–485.

Myasnikova, E. M., Asselain, B., and Yakovlev, A. Yu. (2000), "Spline-Based Estimation of Cure Rates: An Application to the Analysis of Breast Cancer Data," *Mathematical and Computer Modelling*, 32, 217–228.

Nielsen, G., Gill, R., Andersen, P., and Sørensen, T. (1992), "A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models," *Scandinavian Journal of Statistics*, 19, 25–43.

Oakes, D. (1989), "Bivariate Survival Models Induced by Frailties," *Journal of the American Statistical Association*, 84, 487–493.

Peng, Y., and Dear, K. B. G. (2000), "A Nonparametric Mixture Model for Cure Rate Estimation," *Biometrics*, 56, 237–243.

Pepe, M. S., and Fleming, T. R. (1989), "Weighted Kaplan–Meier Statistics: A Class of Distance Tests for Censored Survival Data," *Biometrics*, 45, 497–507.

Press, W., Flannery, B., Teukolsky, S., and Vetterling, W. (1994), *Numerical Recipes in Pascal: The Art of Scientific Computing*, New York: Cambridge University Press.

Rachev, S. T., Wu, C., and Yakovlev, A. Y. (1995), "A Bivariate Limiting Distribution of Tumor Latency Time," *Mathematical Biosciences*, 127, 127–147.

Sposto, R., Sather, H. N., and Baker, S. A. (1992), "A Comparison of Tests of the Difference in the Proportion of Patients Who Are Cured," *Biometrics*, 48, 87–99.

Sy, J. P., and Taylor, J. M. G. (2000), "Estimation in a Cox Proportional Hazards Cure Model," *Biometrics*, 56, 227–236.

——— (2001), "Standard Errors for the Cox Proportional Hazards Cure Model," *Mathematical and Computer Modelling*, 33, 1237–1252.

Taylor, J. M. G. (1995), "Semi-Parametric Estimation in Failure Time Mixture Models," *Biometrics*, 51, 899–907.

Trelford, J., Tsodikov, A. D., and Yakovlev, A. Y. (2001), "Modeling Post-Treatment Development of Cervical Carcinoma: Exophytic or Endophytic—Does It Matter?" *Mathematical and Computer Modelling*, 33, 1439–1443.

Tsodikov, A. (1998a), "A Proportional Hazards Model Taking Account of Long-Term Survivors," *Biometrics*, 54, 1508–1516.

——— (1998b), "Asymptotic Efficiency of a Proportional Hazards Model With Cure," *Statistics & Probability Letters*, 39, 1998, 237–244.

——— (2001), "Estimation of Survival Based on Proportional Hazards When Cure Is a Possibility," *Mathematical and Computer Modelling*, 33, 1227–1236.

——— (2002), "Semiparametric Models of Long- and Short-Term Survival: An Application to the Analysis of Breast Cancer Survival in Utah by Age and Stage," *Statistics in Medicine*, 21, 895–920.

——— (2003), "Semiparametric Models: A Generalized Self-Consistency Approach," submitted to *Journal of the Royal Statistical Society*, Ser. B, 65, 759–774.

Tsodikov, A. D., Asselain, B., Fourquet, A., Hoang, T., and Yakovlev, A. Yu. (1995), "Discrete Strategies of Cancer Post-Treatment Surveillance: Estimation and Optimization Problems," *Biometrics*, 51, 437–447.

Tsodikov, A., Dicello, J., Zaider, M., Zorin, A., and Yakovlev, A. (2001), "Analysis of a Hormesis Effect in the Leukemia Caused Mortality Among Atomic Bomb Survivors," *Human and Ecological Risk Assessment*, 7, 829–847.

Tsodikov, A. D., Hasenclever, D., and Loeffler, M. (1998a), "Regression With Bounded Outcome Score: Evaluation of Power by Bootstrap and Simulation in a Chronic Myelogenous Leukaemia Clinical Trial," *Statistics in Medicine*, 17, 1909–1922.

Tsodikov, A., Loeffler, M., and Yakovlev, A. Yu. (1997), "Assessing the Risk of Secondary Leukemia in Patients Treated for Hodgkin's Disease. A Report From the International Database on Hodgkin's Disease," *Journal of Biological Systems*, 5, 433–444.

——— (1998b), "A Cure Model With Time-Changing Risk Factor: An Application to the Analysis of Secondary Leukemia. A Report From the International Database on Hodgkin's Disease," *Statistics in Medicine*, 17, 27–40.

Tsodikov, A. D., and Müller, W. (1998), "Modeling Carcinogenesis Under a Time-Changing Exposure," *Mathematical Biosciences*, 152, 179–191.

Tucker, S. L. (1999), "Modeling the Probability of Tumor Cure After Fractionated Radiotherapy," in *Mathematical Models in Medical and Health Sciences*, eds. M. A. Horn, G. Simonett, and G. Webb, Nashville: Vanderbilt University Press, pp. 1–15.

Tucker, S. L., and Taylor, J. M. G. (1996), "Improved Models of Tumour Cure," *International Journal of Radiation Biology*, 70, 539–553.

Tucker, S. L., Thames, H. D., and Taylor, J. M. G. (1990), "How Well Is the Probability of Tumor Cure After Fractionated Irradiation Described by Poisson Statistics?" *Radiation Research*, 124, 273–282.

Wassel, J. T., and Moeschberger, M. L. (1993), "A Bivariate Survival Model With Modified Gamma Frailty for Assessing the Impact of Interventions," *Statistics in Medicine*, 12, 241–248.

Yakovlev, A. Yu. (1994), "Letter to the Editor," *Statistics in Medicine*, 13, 983–986.

——— (1996), "Threshold Models of Tumor Recurrence," *Mathematical and Computer Modelling*, 23, 153–164.

Yakovlev, A. Y., Asselain, B., Bardou, V. J., Fourquet, A., Hoang, T., Rochefordiere, A., and Tsodikov, A. D. (1993), "A Simple Stochastic Model of Tumor Recurrence and Its Application to Data on Premenopausal Breast Cancer," in *Biometrie et Analyse de Donnees Spatio-Temporelles*, No. 12, eds. B. Asselain, M. Boniface, C. Duby, C. Lopez, J. P. Masson, and J. Tranchefort, Rennes, France: Société Française de Biométrie, pp. 66–82.

Yakovlev, A. Yu., and Tsodikov, A. D. (1996), *Stochastic Models of Tumor Latency and Their Biostatistical Applications*, Singapore: World Scientific.

Yakovlev, A. Y., Tsodikov, A. D., Boucher, K., and Kerber, R. (1999), "The Shape of the Hazard Function in Breast Carcinoma: Curability of the Disease Revisited," *Cancer*, 85, 1789–1798.

Yamaguchi, K. (1992), "Accelerated Failure-Time Regression Model With a Regression Model of Surviving Fraction: An Analysis of Permanent Employment in Japan," *Journal of the American Statistical Association*, 87, 284–292.

Zaider, M., Zelefsky, M. J., Hanin, L. G., Tsodikov, A. D., Yakovlev, A. Y., and Leibel, S. A. (2001), "A Survival Model for Fractionated Radiotherapy With an Application to Prostate Cancer," *Physics in Medicine and Biology*, 46, 2745–2758.

Zorin, A. V., Tsodikov, A. D., Zharinov, G. M., and Yakovlev, A. Y. (2001), "The Shape of the Hazard Function in Cancer of the Cervix Uteri," *Journal of Biological Systems*, 9, 221–233.

1

Semiparametric Versus Parametric Regression Analysis Based On The Bounded Cumulative Hazard Model: An Application To Breast Cancer Recurrence

Kenneth M. Boucher[a], **Bernard Asselain**[b], **Alexander D. Tsodikov**[a], **and Andrei Y. Yakovlev**[c]

[a]*Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah*
[b]*Biostatistical Unit, Institut Curie, Paris, France*
[c]*Department of Biostatistics and Computational Biology, University of Rochester, Rochester, New York*

Abstract: The bounded cumulative hazard model is a generalization of the proportional hazard model with several distinct advantages:(1) It has superior flexibility, allowing a fit to data where the the proportional hazards assumption does not apply; (2) It has a form that is suitable for semiparametric inference; and (3) It may offer an interpretation in terms of biologically meaningful parameters. In this paper the bounded cumulative hazard model is discussed. Several versions of the model are applied to breast cancer recurrence data from the Curie Institute.

Keywords and phrases: Cure models, Survival models, Bounded cumulative hazard, Semiparametric inference, Parametric models, Breast cancer

1.1 Introduction

In many clinical and epidemiological settings, investigators encounter cause-specific survival curves that tend to level off at a value strictly greater than zero as time increases. This plateau may be taken as an indication of the presence of a proportion of patients for whom the disease under study will never recur. One can consider such patients to be effectively cured. The probability of (biological) cure, variously referred to as the cure rate or the surviving fraction, is defined as an asymptotic value of the survival function $\overline{G}(t)$ as t tends to infinity. Let X be the survival time with cumulative distribution function (c.d.f.) $G(t) = 1 - \bar{G}(t)$.

The existence of a non-zero surviving fraction, p, is determined by the behavior of the hazard function, $\lambda(t)$, by virtue of the equality

$$p = \lim_{t \to \infty} \overline{G}(t) = \exp\left\{ - \int\limits_0^\infty \lambda(u) du \right\}. \tag{1.1}$$

Whenever $p > 0$ (the integral in (1.1) converges), the underlying survival time distribution is said to be improper. Clearly, $\lambda(u) \to 0$ as $u \to \infty$ if $p > 0$ and the limit of $\lambda(u)$ (as $u \to \infty$) exists.

Boag (1949) and later Berkson and Gage (1952) proposed a two-component (binary) mixture model for the analysis of survival data when a proportion of patients is cured. Since then, the binary mixture-based approach has become the dominant one in the literature on cure models (Miller, 1981; Maller and Zhou, 1996). The main idea behind this approach is that any improper survival function can be represented as

$$\bar{G}(t) = \mathrm{E}\left\{ [\bar{G}_0(t)]^M \right\} = p + (1-p)\bar{G}_0(t), \tag{1.2}$$

where M is a binary random variable taking on the values of 0 and 1 with probability p and $1-p$, respectively, with

$$p = \Pr\{X = \infty\},$$

and $\bar{G}_0(t)$ is defined as the survival function for the time to failure conditional upon ultimate failure, i.e.

$$\bar{G}_0(t) = \Pr\{X \geq t | X < \infty\}. \tag{1.3}$$

An alternative, but equally general, representation of an improper survival time distribution can be obtained by assuming that the cumulative hazard $\Lambda(t) = \int_0^t \lambda(t)dt$ has a finite positive limit, say θ, as t tends to infinity. In this case, one can write

$$\bar{G}(t) = e^{-\theta F(t)}, \quad \theta > 0, \quad t \geq 0, \tag{1.4}$$

where $F(t) = \Lambda(t)/\theta$ is the c.d.f. of some non-negative random variable such that $F(0) = 0$. In what follows, we will call the model given by (1.4) the bounded cumulative hazard (BCH) model.

Clearly, estimating the proportion of cured patients may have important medical implications. In addition, clinical covariates may exert dissimilar effects on the probability of cure and the timing of tumor relapse or other events of interest. There are at least two advantages of the BCH model: (1) they enrich our ability to interpret survival analysis in terms of characteristics that have a clear biomedical meaning; (2) they lead to more general regression models, thereby extending our ability to describe actual data. It took a long time to realize these virtues of the BCH model. The first move in this direction was due to Haybittle (1959, 1965). The author proceeded from the observation that in some clinical data on cancer survival, an actuarial estimate of the hazard function tends to decrease exponentially with time. If the same property holds for the true hazard, expression (1.4) assumes the form:

$$\bar{G}(t) = e^{-\theta(1-e^{-\alpha t})}, \quad \alpha > 0. \tag{1.5}$$

A comprehensive treatment of this model was given by Cantor and Shuster (1992). Clearly, the Gompertz-like model given by formula (1.5) is a special case of formula (1.4) with the function $F(t)$ specified by an exponential c.d.f. with parameter α.

The representation (1.4) for the BCH model was first introduced in the paper of Yakovlev et al. (1993) and discussed later as an alternative to the binary mixture model by Yakovlev (1994). Interestingly enough, Yakovlev et al. (1993) proceeded from purely biological considerations; the idea of imposing a constraint on the behavior of the hazard function was introduced much later in Tsodikov et al. (1998). In fact, the authors proposed a simple mechanistically motivated model of tumor recurrence yielding an improper survival time distribution. Under this model, the probability of tumor cure is defined as the probability of no clonogenic tumor cells surviving by the end of treatment. A comprehensive account of the BCH model and associated statistical methods is given by Tsodikov, Ibrahim, and Yakovlev (2003).

In this paper, we discuss methods of semiparametric inference based on the BCH model that have been recently developed. We then illustrate the method by applying several (semiparametric and parametric) versions of the BCH model to breast cancer recurrence data from the Curie Institute.

1.2 BCH Regression Models

1.2.1 Mixture Models and Generalizations

In this subsection we discuss how certain BCM models, and in particular, the double proportional hazards model, arise naturally within a more general class of regression models, called Nonlinear Transformation Models.

Perhaps the simplest bounded cumulative hazard model is obtained by making the assumption that the cumulative hazard is bounded in a proportional hazards (PH) model. In this way we obtain the so-called improper PH model

$$\bar{G}(t|z) = \exp\{-\theta(\beta, z)F(t)\}, \tag{1.6}$$

where β is a vector of regression coefficients, and z is a vector of covariates, and $-\theta(\beta, z)$ is a fixed (known) function relating β to z. A general class of regression models, Nonlinear Transformation Models (NTM), was proposed in Tsodikov (2002,2003), in the context of semiparametric models:

$$\bar{G}(t|z) = \gamma(\bar{F}(t)|\beta, z), \tag{1.7}$$

where $\gamma(x|\beta, z)$ is some parametrically specified cumulative distribution function in x with support on $[0, 1]$. In the examples we assume γ is parametrized through a set of parameters/predictors $\theta, \eta, ...$, where each predictor is further parametrized using (generally different) sets of regression coefficients $\beta_1, \beta_2, ...,,$ so that $\theta = \exp\{\beta_1' z\}, \eta = \exp\{\beta_2' z\}$. Linear transformation models considered by Cheng et al. (1995) represent a subclass of NTM. In order to make (1.7) a BCH model, the following assumptions are made to enforce the limit $\bar{G}(t|\beta, z) \to p > 0$:

$$\gamma(0|\beta, z) > 0, \ \lim_{t\to\infty} \bar{F}(t) = 0. \tag{1.8}$$

The restriction $\bar{F}$(last failure) $= 0$ proposed by Taylor (1995) in the context of the two-component mixture model removes the over-parameterization of the description of the baseline cure rate through F and h. Following the restriction, the estimate $\hat{\bar{F}}$ is assumed to satisfy $\hat{\bar{F}}$(last failure) $= 0$. Moreover, it is necessary for separation of the long- and short-term covariate effects on survival.

Alternatively, a BCH model can be formulated as a two-stage model. First, an unobservable random variable M is postulated with distribution $p(m|z)$ that depends on covariates. Second, the observed survival function is formed as

$$\bar{G}(t|\beta, z) = \mathrm{E}\left[\bar{F}(t)^M \middle| z\right], \tag{1.9}$$

where the expectation is taken with respect to M given z, and the distribution of M depends on β and z . The model (1.9) represents a generalized PH frailty model, which we simply call the PH mixture model (Tsodikov,2002; 2003), where M is assumed to be an arbitrary non-negative random variable. In this context, M can be interpreted as a missing covariate in a PH model. Since all models considered previously in this paper are particular cases of (1.9), an estimation procedure designed for the PH mixture model represents a universal tool for statistical inference with such models.

The discrete mixture model (i.e. (1.9) with a discrete random variable M) can be linked to the NTM class. Observe that the probability generating function of a discrete random variable M

$$\varphi_M(x) = \sum_{m=0}^{\infty} p_m x^m$$

is a distribution function in x with the support [0,1]. Indeed, since p_m is non-negative, $\varphi_M(x)$ is increasing in x. Also, $\varphi_M(1) = \sum_{m=0}^{\infty} p_m = 1$. If M is discrete and depends on covariates, (1.9) can be rewritten as an NTM with

$$\gamma(x|z) = \varphi_M(x|z).$$

With the discrete mixture model, the probability of cure is given by $\gamma(0|z) = \varphi_M(0|\beta, z) = p_0(\beta, z)$.

We can now represent both mixture and BCH models discussed earlier in this paper as members of the Mixture - NTM model family. Consider an extension of the improper PH model allowing for dissimilar covariate effects on long- and short-term survival. To construct an extended hazard model (the term was introduced by Etezadi-Amoli and Ciampi (1987)), we employ the fact that $\bar{F}$ is a survival function. Incorporating covariates into $\bar{F}$, we can add a short-term effect to the improper PH model. The class of extended hazard models

$$\bar{G}(t|\beta, z) = \exp\left\{-\theta(\beta_1, z)[1 - \tilde{\gamma}(\bar{F}(t)|\beta_2, z)]\right\}, \qquad (1.10)$$

where $\tilde{\gamma}$ is an NTM, $\beta = (\beta_1, \beta_2)$, $\beta_i = (\beta_{0i}, \beta_{1i}, ...), i = 1, 2$ was introduced in Tsodikov (2002). If $\tilde{\gamma}$ is a mixture model itself, then (1.10) is also a mixture model with $\gamma = \exp\{-\theta[1 - \tilde{\gamma}]\}$. The mixing variable M that generates the family (1.10) has a well defined structure of the compound Poisson variable

$$M = \sum_{k=1}^{\nu} \xi_k, \qquad (1.11)$$

where ν is a Poisson random variable, and ξ_k are i.i.d. copies of a random variable ξ with Laplace transform $\tilde{\gamma}(e^{-\lambda})$. The binary distribution for ν gives rise to the two-component mixture class of models

$$\bar{G}(t|\beta, z) = p(\beta_1, z) + \bar{p}(\beta_1, z)\tilde{\gamma}[\bar{F}(t)|\beta_2, z], \bar{p} = 1 - p. \qquad (1.12)$$

Kuk and Chen (1992) proposed a regression model in the form of (1.12) with p being a logistic regression, and $\tilde{\gamma}$ a proportional hazards regression. This model was further studied by Sy and Taylor (2000) and Peng and Dear (2000). The Poisson distribution for ν leads to the bounded hazard family of mixture models. For example, let ξ be degenerate (nonrandom): $\Pr(\xi = \eta(z)) = 1$. Then $M = \eta(\beta_2, z)\nu$, where ν is Poisson with expectation $\theta(z)$. With the parametrization $\theta(\beta_1, z) = \exp(\beta_1' z)$ and $\eta(\beta_2, z) = \exp(\beta_2' z)$, we have the PHPH model

$$G(t|\beta, z) = \exp\left[-\exp(\beta_1' z)\left\{1 - \bar{F}^{\exp(\beta_2' z)}\right\}\right], \qquad (1.13)$$

The model (1.13) was proposed by Broët et al. (2001) in the context of two sample score tests for long- and short-term covariate effects. An intercept term (log(cure rate)) is included in β_1 but not β_2. With β_{20} set to zero, β_{10} may be interpreted as the baseline log cure rate. With the exponential parametrization of predictors (1.13), the baseline survival function takes the form $G_b(t|\beta_0, z) = \exp\left[-\exp(\beta_{10} z)\left\{1 - \tilde{\gamma}(\bar{F}|\mathbf{0})\right\}\right]$, where $\mathbf{0}$=(0, 0, ...) and $\beta_0 = ((\beta_{10}, 0, 0, ...,), \mathbf{0})$. As in the Cox model (Cox, 1972), the regression coefficient β_{ij}, $i = 1, 2$, $j = 1, 2, ...$ correspond to the relative effects of the covariate z_{ij} on the long- or short-term predictor, $i = 1, 2$, respectively.

At times an extension of (1.7) may be convenient. For brevity, we describe the extension only for the case we need later. Let $\bar{F}_1$ and $\bar{F}_2$ be two (usually parametrically specified) survival functions. Consider the model

$$G(t|\beta, z) = \exp\left[-\exp(\beta_1' z)\left\{1 - \bar{F}_1\right\} - \exp(\beta_2' z)\left\{1 - \bar{F}_2\right\}\right]. \qquad (1.14)$$

Models of this type are called competing risks models, since F_1 and F_2 may clearly be interpreted as the c.d.f's for competing causes of failure. If we stipulate that the means μ_1 and μ_2 of the corresponding c.d.f's F_1 and F_2 satisfy $\mu_1 < \mu_2$, then θ_1 and θ_2 may be interpreted in terms short-term and long-term effects, respectively. In this way we have can form a very different model from the PHPH model with a similar interpretation. Note that if we impose the pair of restrictions $\lim_{t\to\infty} \bar{F}_1(t) = 0$, and $\lim_{t\to\infty} \bar{F}_2(t) = 0$, the cumulative hazard is bounded, so that (1.14) is in the class of models (1.4), with $\theta = \theta_1 + \theta_2$.

1.2.2 An EM-based estimation procedure for the PH mixture model

The issue of potentially unlimited dimension has been the most critical deterrent to the use of maximum likelihood estimation (MLE) in semiparametric regression models. Methods based on partial likelihood are specific to the proportional hazards model, and do not extend to other models. The Newton-Raphson procedure requires taking inverse of the information matrix, which gets computationally prohibitive and unstable with increasing dimension. Characterizing the future directions in survival analysis in their recent editorial in *Biometrics*, Fleming and Lin (2000) pointed out that "it would be highly useful to develop efficient and reliable numerical algorithms for the semiparametric estimation...".

Recently, Tsodikov (2003) generalized the EM algorithm for the frailty model into a universal 'distribution-free' procedure applicable to the NTM class (1.7). This family of algorithms (Quasi-EM) is a subclass of the so-called MM algorithms based on surrogate objective functions (Lange, Hunter, and Yang, 2000). Broadly defined, an MM algorithm substitutes a computationally simpler surrogate objective function for the target function on each step of the procedure (similar to the E-step of the EM). Maximizing the surrogate objective function drives the target function in the correct direction. Thus, a difficult maximization procedure is replaced by a simpler one. The idea of the QEM approach is to obtain a surrogate objective function that does not depend on the entire distribution of the missing data, that is, the derivation of a 'distribution-free' E-step for the PH mixture model. The following statement is the key result.

Proposition 1.2.1 *(Tsodikov, 2003) Let τ be the observed event time, and c the observed censoring indicator (c = 1 for failures, and c = 0 otherwise).*

Under the PH mixture model

$$G(t|\cdot) = \gamma[\bar{F}(t)|\cdot] = E\{[\bar{F}(t)]^{M(\cdot)}\},$$

and if $F'(t) > 0$, the conditional expectation of M, given the observed event, is given by

$$E\{M(\cdot)|\cdot, \tau, c\} = \Theta[\bar{F}(\tau)|\cdot, c],$$

where

$$\Theta[x|\cdot, c] = c + x\frac{\gamma^{(c+1)}(x|\cdot)}{\gamma^{(c)}(x|\cdot)},$$

and $\gamma^{(i)}(x|\cdot) = \partial^i\gamma(x|\cdot)/\partial x^i$, $i = 1, 2$ and $\gamma^{(0)}(x|\cdot) = \gamma(x|\cdot)$.

■

The above statement indicates that the E-step can be constructed using the first two derivatives of the NTM-generating function γ without any knowledge or even existence of the mixing random variable M. Specification of the algorithm for a particular model requires evaluation of Θ for that model.

Cure models require a correction of Θ at the last observation (see 1.16) below). Introduce a set of times t_i, $i = 1, \ldots, n$, arranged in increasing order, where $t_{n+1} = \infty$. Associated with each t_i is a set of individuals $\mathcal{D}_i$ with covariates z_{ij}, $j \in \mathcal{D}_i$ who fail at t_i, and a similar set of individuals $\mathcal{C}_i$ with covariates z_{ij}, $j \in \mathcal{C}_i$ who are censored at t_i.

According to the nonparametric maximum likelihood method, the model can be fitted by maximizing the generalized loglikelihood ℓ with respect to the regression parameters β and unspecified survival function $\bar{F}$ (or the corresponding distribution function F). An argument similar to that adopted by Kalbfleisch and Prentice (1980) can be used to verify that ℓ is maximized by a step-function F with steps at the times of failures. Let τ_k, $k = 1, \ldots, K$ be the time points in ascending order where failures occur ($\mathcal{D}$ is not empty). Denote by r_k the rank of τ_k in the set $\{t_i\}$. The ranks r_k locate the points τ_k on the net $\{t_i\}$: $\tau_k = t_{r_k}$. By definition we set $\tau_{K+1} = n + 1$ and $\tau_0 = 0$. For any $A(t)$ let $A_i = A(t_i)$.

Differentiating the generalized loglikelihood with respect to $\Delta H_k = H_k - H_{k-1}$, $k = 1, \ldots, r_{K-1}$, we obtain the score equations for F in the form

$$\Delta H_k = \frac{D_k}{\sum_{ij \in \mathcal{R}_k} \Theta(\bar{F}_i|\beta, z_{ij})}, \tag{1.15}$$

where $\mathcal{R}_k = \bigcup_{i=k}^{n} \mathcal{D}_i \cup \mathcal{C}_i$ is the risk set at t_k, D_k is the number of failures at

t_k, and

$$\Theta(\bar{F}_i|\beta, z_{ij}) = \begin{cases} \frac{\gamma'(\bar{F}_i|\beta,z_{ij})}{\gamma(\bar{F}_i|\beta,z_{ij})}\bar{F}_i, & j \in \mathcal{C}_i,\ i \leq r_K - 1 \\ 1 + \frac{\gamma''(\bar{F}_i|\beta,z_{ij})}{\gamma'(\bar{F}_i|\beta,z_{ij})}\bar{F}_i, & j \in \mathcal{D}_i,\ i \leq r_{K-1} \\ \frac{\gamma'(\bar{F}_{r_{K-1}}|\beta,z_{ij})}{\gamma(\bar{F}_{r_{K-1}}|\beta,z_{ij})-\gamma(0|\beta,z_{ij})}\bar{F}_{r_{K-1}}, & j \in \mathcal{D}_{r_K},\ i = r_K \\ 0, & j \in \mathcal{C}_i,\ i \geq r_K. \end{cases} \tag{1.16}$$

The fact that the contributions of the observations ij to the likelihood as well as Θ at, or after the last failure ($i \geq r_K$), are different from their counterparts before the last failure ($i \leq r_K - 1$) is due to the restriction $\bar{F}_i = 0$, $i = r_K, \ldots, n$ representing the fact that $\bar{F}$ is a proper survival function. The above distinction is not made if the model is a general NTM with unrestricted F.

It is interesting to note that the score equations have the form of the Nelson-Aalen-Breslow estimator for the PH model with the usual predictor replaced by Θ. Given β, iterations with respect to F can be carried out as follows:

qE: With kth-iteration $\bar{F}^{(k)}$, compute $\Theta^{(k)}$ for each subject.

MA: Update F using the Nelson-Aalen-Breslow estimator (1.15),

It can be shown (Tsodikov, 2003) that if $\Theta(x, \cdot)$ is a nondecreasing function of x, or if γ is a PH mixture model (a stronger assumption), then each iteration described above improves the likelihood. Also, this assumption makes the above procedure a member of the MM family (Lange, Hunter, and Yang, 2000), and the convergence properties follow from the general MM theory.

There are many ways to build a particular model fitting algorithm based on the principles above, and this depends on how the maximization with respect to β is incorporated into the procedure. One possibility is to maximize the profile likelihood

$$\ell_{pr}(\beta) = \ell(\beta, \hat{\bar{F}}(\beta)), \tag{1.17}$$

with respect to β, where $\hat{\bar{F}}(\beta)$ is determined at each step of maximization by iterating the steps (qE) and (MA) until convergence. This method was used for the semiparametric model used in the analysis presented below.

1.3 Analysis of Data from the Curie Institute

Two bounded cumulative hazards models, the double proportional hazards (1.13) and the competing risks model (1.14) were applied to data from women diagnosed with primary breast carcinoma between 1981 and 1991 at the Curie

Institute. The endpoint of interest was recurrence in the ipsilateral breast. Both models admit interpretation in terms of long and short term effects, so sensitivity to model specification could be explored. In addition we used three versions of the PHPH model, differing in the way $\bar{F}$ was specified, to explore sensitivity to specification of $\bar{F}$.

The general treatment policy for the Curie Institute data was aimed towards breast conservation through a combined application of radiotherapy and limited surgery. A detailed description of subcohorts of these patients is given in Fourquet et al (1989). The strategy of the analysis was to choose important predictors using backward elimination. Women with missing tumor size, grade, nodal status, estrogen receptor status, progesterone receptor status, or treatment information were excluded from the analysis. Women for whom grade was coded as 'not done' were included however. After these exclusions, a sample of 6232 women out of an initial total of 9899 was available for analysis. The vast majority (3278, or 89%) of the excluded patients had either missing estrogen or progesterone receptor status, or both.

The initial model from which backward elimination proceeded included both long- and short-term effects for all variables. The variables included were: primary tumor size at detection, estrogen and progesterone receptor status, clinical axillary lymph node status (nodal involvement), age, and tumor size, histological grade, and treatment. Estrogen receptor status (ER), progesterone receptor status (PgR) and nodal involvement were taken to be indicator variables. Treatment was taken to be a categorical variable with four levels: radiotherapy alone (RT), mastectomy alone, tumorectomy plus radiotherapy, and mastectomy plus radiotherapy. Age (in years) at diagnosis and primary tumor size (in millimeters) were considered as continuous covariates in this analysis. Histological grading (Scarff Bloom) was taken as a variable with three categories (I, IIa and IIb combined, and III). Tumor grade was highly correlated with stage, and, in combination with the other covariates, was chosen preferentially over stage as predictive of local recurrence in a preliminary analysis with all covariates. Each of the predictors included in the backward elimination was highly significant in univariate analysis.

The three PHPH models differed in the way in which $\bar{F}$ was specified. Our most flexible model used a semiparametric specification of $\bar{F}$. We refer to this model as the 'semiparametric PHPH model'. Our intermediate form specified that $\bar{F}$ be a linear spline. For our spline models we use fixed, equally spaced knots between $t = 0$ and $t = t_{max}$, where t_{max} is the last failure time. $\bar{F}$ is parametrized by the values at the (ten) knots, subject to the conditions that $\bar{F}$ is monotone decreasing, $\bar{F}(0) = 1$, and $\bar{F}(t_{max}) = 0$. Linear interpolation is used between the knots to ensure continuity. We refer to this model as the 'spline PHPH model'. The most restrictive parametric form specified that $\bar{F}$ have a Weibull distribution. This model uses only two parameters, (a median and shape parameter), to specify $\bar{F}$. We refer to this model as 'Weibull PHPH model'. Weibull distributions were also used for both $\bar{F}_1$ and $\bar{F}_2$ of the competing risks model. A semiparametric or spline-based form of the competing risks model was not used, as it is difficult to retain identifiability with such a model.

Estimates for semiparametric PHPH model were obtained by maximizing the profile likelihood (1.17) using the Powell algorithm (Himmelblau, 1972), as described in Section 1.2.2. We fit the parametric and spline PHPH models, as well as the competing risks model, by backward elimination as well. For these models, the complete likelihood was maximized via the Powell algorithm. Variables with several levels were considered in blocks. Table 1.1 gives the significance levels of variables at the step of removal or inclusion in the final model, calculated using the likelihood ratio test. The cycle in which predictors were removed is given for all variables not in the final model. The significance level $\alpha = 0.05$ was chosen for inclusion in the final model. Although predictors are removed in a different order, in the end the final list of predictors selected for the semiparametric PHPH and Weibull-based PHPH models were exactly the same. One more short-term predictor (size) was selected as significant for for the spline PHPH model. The competing risks model contained the additional short-term effect of age (but no size parameters). We note that in each case, there are highly significant short-term effects, so that the proportional hazards assumption is easily rejected.

Parameter estimates for each of the models are provided in Table 1.2. The parametrization for the PHPH and competing risks models are given by (1.13) and (1.14) respectively. Note that the cure rate estimates, although comparable to each other, apply to the extrapolated 'baseline' subject with age equal to zero and tumor size equal to zero. For the cure rate model, the median for one component, interpreted as long term, was 123 years, while the median for the other component, labeled short term, was 23 years.

The models were generally in very good agreement. For all three PHPH models and the competing risks model, positive progesterone receptor status and older age all significantly increased the probability of 'local cure', while lack of nodal involvement, positive estrogen receptor status, and low histological grade significantly increased the mean time to local recurrence for uncured patients. Interestingly, tumor size was chosen as (marginally) significant only for the spline PHPH model. For the spline PHPH model, there was a significant decrease in the mean time to recurrence with increasing primary tumor size. For the competing risk model, older age had a significant short term benefit as well as a long term benefit. The cure rate parameters were in good agreement for all the PHPH models, but the cure rate parameter for the competing risks model varied somewhat, perhaps because of the quite different model specification.

Treatment had both a highly significant long term and short term effect. Compared to the 'baseline' group of patients receiving radiotherapy alone, every

Effect	**Predictor**	**Model**			
		Semi. PHPH	Spline PHPH	Weibull PHPH	Comp. Risks
Long Term	Nodes	1 (0.92)	1 (0.78)	2 (0.88)	4 (0.18)
	PrR	* (0.02)	* (0.02)	* (0.02)	* (0.01)
	ER	4 (0.23)	6 (0.09)	5 (0.20)	2 (0.94)
	Histology	5 (0.16)	5 (0.14)	6 (0.15)	1 (0.90)
	Treatment	* (E-13)	* (E-12)	* (E-15)	* (E-13)
	Age	* (E-13)	* (E-13)	* (E-13)	* (1E-9)
	Size	2 (0.86)	3 (0.51)	1 (0.99)	3 (0.69)
Short Term	Nodes	* (3E-4)	* (2E-4)	* (1E-6)	* (7E-7)
	PgR	3 (0.22)	4 (0.19)	4 (0.18)	5 (0.14)
	ER	* (5E-9)	* (4E-9)	* (E-10)	* (1E-9)
	Histology	* (1E-8)	* (8E-9)	* (2E-9)	* (7E-9)
	Treatment	* (9E-5)	* (5E-5)	* (6E-7)	* (1E-6)
	Age	6 (0.12)	2 (0.58)	3 (0.25)	* (8E-4)
	Size	7 (0.10)	* (0.04)	7 (0.07)	6 (0.09)

Table 1.1: Step at which predictors were removed, and significance level (in parentheses) for the three PHPH models and the competing risks model fitted to breast cancer data using backward elimination. Predictors that were not removed are indicated by *. The significance level applies either to the removal step (for predictors that were eliminated by backward selection) or the final model (for all the other predictors).

other group of patients had a greater chance of local cure for all the models, with those receiving mastectomy (with or without RT) having the best chance of local cure. Curiously, according to the all the PHPH models, compared to patients receiving only radiotherapy, patients receiving mastectomy plus radiotherapy or mastectomy alone, and who have not been locally cured, had smaller mean time to recurrence. On the other hand, patients who received lumpectomy plus radiotherapy were less likely to have been cured locally than those receiving mastectomy or mastectomy plus radiotherapy, but the mean time to recurrence was increased. For the competing risk model mastectomy, mastectomy plus RT, and lumpectomy + RT all had comparable beneficial effect on time to recurrence compared to RT alone.

1.4 Discussion

Bounded cumulative hazards regression models have advantages over other models: they are readily interpretable in terms of characteristics with biomedical meaning, and they have more flexibility than more common modeling tools, and can be fit using semiparametic methods. These advantages have been illustrated by our application to the Curie Institute data, presented in the preceding section. Even though our results appear to be in agreement with the literature, our approach, which includes both long and short term effects, provides more information than the typical proportional hazards modeling. From our model we can see that estrogen receptor status, nodal involvement, and histological grade appear to largely affect the timing of local recurrence, rather than the probability of local cure. Age, on the other hand, has a strong effect on the probability of local cure, and treatment on both the probability of local cure as well as the timing of recurrence.

The results of our application of the PHPH and competing risks models to the Curie Institute data suggest that the long and short terms effects are largely robust to changes in model specification (from PHPH to competing risks model) or changes in specification of time to failure distributions. In addition, our results appear to be largely consistent with the breast cancer literature: decreased probability of local relapse is associated with older age at onset, negative nodes, positive estrogen receptor status, and low histological grade, particularly after breast conserving therapy. (See, for example, McReady et al. (1996), who analyzed breast cancer treated with lumpectomy alone). Young age is often regarded as an important risk factor not only for local recurrence, but distant metastasis and shorter overall survival as well (Elkhuizen et al., 1998). Irradiation of the breast in patients receiving lumpectomy is regarded as reducing risk of recurrence, except, perhaps, for patients at very low risk

			Model			
Effect	**Predictor**	**Level**	Semi.	Spline	Weibull	Comp.
			PHPH	PHPH	PHPH	Risks
Long Term	PgR	Negative	—	—	—	—
		Positive	-0.16	-0.16	-0.16	-0.25
	Treatment	RT	—	—	—	—
		Mastectomy	-1.29	-1.27	-1.28	-1.26
		RT+Lump.	-0.40	-0.38	-0.38	-0.61
		RT+Mast.	-1.57	-1.57	-1.53	-1.55
	Age (yrs.)	(continuous)	-0.023	-0.023	-0.023	-0.026
	Cure Rate		0.064	0.064	0.068	0.043
Short Term	Nodes	Negative	—	—	—	—
		Positive	0.47	0.42	0.50	0.66
	ER	Negative	—	—	—	—
		Positive	-0.60	-0.60	-0.64	-0.81
	Histology	Not done	—	—	—	—
		I	-0.57	-0.53	-0.64	-1.11
		II	-0.17	-0.13	-0.20	0.15
		III	0.29	0.34	0.25	0.44
	Treatment	RT	—	—	—	—
		Mastectomy	0.36	0.39	0.37	-0.74
		RT+Lump.	-0.47	-0.39	-0.52	-0.87
		RT+Mast.	0.67	0.70	0.63	-0.90
	Age (yrs.)	(continuous)	***	***	***	-0.020
	Size (mm.)	(continuous)	***	0.0049	***	***

Table 1.2: Parameter estimates for each of the three versions of the PHPH model, and the competing risks model, fit to breast cancer recurrence data from the Curie Institute. The baseline group is indicated by —. Components that are absent from the model are indicated by ***.

(Clark et al., 1996).

Association of small tumor size with decreased risk of local recurrence has been observed in the literature. (See, for example, McGreedy, 1996, or Asselain et al., 1996). We do not see much effect of tumor size, as size was absent from two out of three of our final PHPH models, was only marginally significant in the short-term component of the other models, and never appeared in the long-term component of the models. A further exploration of this is discussed below.

A tumor recurrence data set containing a subset of 877 patients from the Curie Institute data was analyzed by Asselain et al. (1996) using the accelerated failure time model to describe short-term covariate effects. Although recurrence in both the ipsilateral and contralateral breast were analyzed together, the authors state that analysis of ipsilateral recurrence alone gave similar results. The model incorporated components that could be interpreted in terms of long-term effects (mean number of surviving clonogens) and short-term effects (time of progression). A more limited set of covariates (age, tumor size, nodal involvement, and treatment) was considered in this analysis. After backward elimination, treatment, the short-term effect of tumor size, and the long-term effects of both age and tumor size. In accordance with the analysis presented in the previous section, age appeared to have a pronounced long-term effect, but little effect on the time of tumor recurrence. The authors indicate that the long-term effect of age appeared to be the predominant one.

The apparent lack of significance of tumor size in the current analysis appears to contradict the results of the Asselain et al. (1996), but it is possible that this is due to different lists of predictors. To check whether the inclusion of receptor status in the current analysis is responsible for the discrepancy, we removed estrogen and progesterone receptor status from the list of predictors and fit the semiparametric PHPH model to the data by backwards elimination. There is a significant short-term effect of tumor size when this is done ($p=0.026$). There is, in addition, a significant short-term effect of age ($p = 0.027$), in addition to the long-term effect that was present in the extended model. The other selected predictors are the same. We see that even after a more similar list of predictors are included, there remains some slight discrepancy between the effects of tumor size. We speculate that this discrepancy may be caused by the inclusion of more detail on radiotherapy in the earlier analysis. Shorter follow-up analysis and smaller sample size may also contribute to the discrepancy.

Finally, we note that if a different endpoint is analyzed, such as cause-specific survival or time to metastatic spread, the effect of age at onset may be different. For example, in an analysis of survival in women diagnosed with localized breast cancer in Utah using a similar model, with age and stage as predictors, Tsodikov (2002) found that the long term effect of age at diagnosis was not significant, while the short term effect remained significant, at least for younger patients.

Acknowledgment

This research was supported by NIA/NIH Grant RO1 AG14650-01A2, NCI/NIH U01 CA88177-01, NIH/NCI 1 UO1 CA97414-01 and a Department of Defense grant DAMD17-03-1-0034.

References

1. Asselain, B., Fourquet, A., Hoang, T., Tsodikov, A.D. and Yakovlev, A.Y. (1996). A parametric regression model of tumor recurrence: An application to the analysis of clinical breast cancer data, *Statistics & Probability Letters*, **29**, 271–278.

2. Berkson, J. and Gage, R.P. (1952). Survival curves for cancer patients following treatment, *J. Amer. Statist. Ass.*, **47**, 501–515.

3. Boag, J.M. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy, *J. Roy. Statist. Soc., Ser. B*, **11**, 15–44.

4. Broët, P., Rycke, Y.D., Tubert-Bitter, P., Lellouch, J., Asselain, B., and Moreau, T. (2001). A semi-parametric approach for the two-sample comparison of survival times with long-term survivors, *Biometrics*, **57**, 844–852.

5. Cantor, A.B. and Shuster, J.J. (1992), Parametric versus nonparametric methods for estimating cure rates based on censored survival data, *Statistics in Medicine*, **11**, 931–937.

6. Cheng, S., L. Wei, and Z. Ying (1995). Analysis of transformation models with censored data, *Biometrika*,**82**, 835–845.

7. Clark, R.M., Whelan, T., Levin, M., Roberts, R., Willan, A., McCulloch, P., Lipa, M., Wilkinson, R.H., and Mahoney, L.J. (1996). Randomized clinical trial of breast irradiation following lumpectomy and axillary dissection for node-negative breast cancer: an update. Ontario Clinical Oncology Group, *Journal Natl. Cancer Institute*, **88**(22), 1659–64.

8. Cox, D. R. (1972), Regression Models and Life Tables, (with discussion), *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.

9. Elkhuizen, P.H., van de Vijver, M.J., Hermans, J., Zonderland, H.M., van de Velde, C.J., and Leer, J.W. (1998). Local recurrence after breast-conserving therapy for breast cancer: high incidence in young patients and association with poor survival, *Int. J. Rad. Oncol. Biol. Phys*, **40**(4) 859-67.

10. Etezadi-Amoli, L. and Ciampi, A. (1987). Extended hazard regression for censored survival data with covariates: A spline approximation for the baseline hazard function, *Biometrics* **43**, 181–192.

11. Fleming, T. and Lin, D. (2000). Survival analysis in clinical trials: Past developments and future directions, *Biometrics* **56**, 971–983.

12. Fourquet, A.F., Campana, F., Zafrani, V, Mosseri, P., Vielh, P., Durand, J.-C., and Vilcoq, J.R. (1989). Prognostic factors of breast recurrence in the conservative management of early breast cancer: A 25 year follow-up. *Int. J. Radiat. Onc. Biol. Phys.* **17**, 719–725.

13. Haybittle, J.L. (1959). The estimation of the proportion of patients cured after treatment for cancer of the breast. *Br. J. Radiol.*, **32**, 725-733.

14. Haybittle, J.L. (1965). A two-parameter model for the survival curve of treated cancer patients. *J. Amer. Statist. Assoc.*, **60**, 16–26.

15. Himmelblau, D.M. (1972). *Applied Nonlinear Programming*, McGraw-Hill, Austin.

16. Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data.* John Wiley & Sons, New York.

17. Kuk, A.Y.C., and Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression, *Biometrika*, **79**, 531–541.

18. Lange, K. and Hunter, D.R., and Yang, I.(2000). Optimization transfer using surrogate objective functions (with discussion), *Journal of Computational and Graphical Statistics*, **9** 1–59.

19. McCready, D.R., Hanna, W., Kahn, H., Chapman, J.A., Wall, J., Fish, E.B., and Lickley, H.L. (1996). Factors associated with local breast cancer recurrence after lumpectomy alone, *Ann. Surg. Oncol.* **3**(4), 358-66.

20. Maller, R.A. and Zhou, S. (1996). *Survival Analysis with Long-Term Survivors*, Wiley, Chichester.

21. Miller, R.G. (1981). *Survival Analysis*, John Wiley, New York.

22. Nielsen, G., Gill, R., Andersen, P. and Sorensen, T. (1992). A counting process approach to maximum likelihood estimation in frailty models, *Scand. J. Statist.*, **19**, 25–43.

23. Peng, Y., and Dear, K.B.G. (2000). A nonparametric mixture model for cure rate estimation, *Biometrics*, **56**, 237–243.

24. Sy, J.P., and Taylor, J.M.G. (2000). Estimation in a Cox proportional hazards cure model, *Biometrics*, **56**, 227–236.

25. Taylor, J.M.G. (1995). Semi-parametric estimation in failure time mixture models, *Biometrics*, **51**, 899–907.

26. Tsodikov, A. (1998). Asymptotic efficiency of a proportional hazards model with cure. *Stat. Probab. Letters*, **39**, 1998, 237–244.

27. Tsodikov, A. (2002). Semiparametric models of long- and short-term survival: An application to the analysis of breast cancer survival in Utah by age and stage, *Statistics in Medicine*, **21**, 895–920.

28. Tsodikov, A. (2003). Semiparametric models: A generalized self-consistency approach. *J. Roy. Statist. Soc., Ser. B*, in press.

29. Tsodikov, A.D., Ibrahim, J.G., and Yakovlev, A.Y. (2003). Estimating Cure Rates from Survival Data: An Alternative to Two-Component Mixture Models, *J. Amer. Statist. Assoc.*, in press.

30. Yakovlev, A.Yu. (1994). Letter to the Editor. *Statistics in Medicine* **13**, 983–986.

31. Yakovlev, A.Y., Asselain, B., Bardou, V.J., Fourquet, A., Hoang, T., Rochefordiere, A., and Tsodikov, A.D. (1993). A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer, In *Biometrie et Analyse de Donnees Spatio-Temporelles* No 12, (Eds. B. Asselain, M. Boniface, C. Duby, C. Lopez, J.P. Masson, and J. Tranchefort), Société Française de Biométrie, ENSA Rennes, France, pp. 66–82.